Theory of Probability

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

Theory of Probability

A Critical Introductory Treatment

Bruno de Finetti

Translated by Antonio Machí and Adrian Smith

WILEY

This edition first published 2017

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

This new publication is Volume I and Volume II combined.

Previous edition first published in 1970 © Giulio Einaudi, *Teoria Delle Probabilita* – Bruno de Finetti
Republished by John Wiley & Sons Ltd in 1975 © *Theory of Probability* – Bruno de Finetti

Library of Congress Cataloging-in-Publication Data

Names: de Finetti, Bruno, author. | Machí, Antonio, translator. | Smith, Adrian F. M., translator. | de Finetti, Bruno. Teoria delle probabilità. English.
Title: Theory of probability : a critical introductory treatment / Bruno de Finetti ; translated by Antonio Machi, Adrian Smith.
Description: Chichester, UK ; Hoboken, NJ : John Wiley & Sons, 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2016031568 (print) | LCCN 2016049185 (ebook) | ISBN 9781119286370 (cloth : alk. paper) | ISBN 9781119286349 (Adobe PDF) | ISBN 9781119286295 (ePub)
Subjects: LCSH: Probabilities.
Classification: LCC QA273.A5 D4 2017 (print) | LCC QA273.A5 (ebook) | DDC 519.2–dc23
LC record available at https://lccn.loc.gov/2016031568

A catalogue record for this book is available from the British Library.

Cover Design: Wiley

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

This work is dedicated to my colleague Beniamino Segre who about twenty years ago pressed me to write it as a necessary document for clarifying one point of view in its entirety

[1970]

Contents

Foreword

It is an honour to be asked to write a foreword to this book, for I believe that it is a book destined ultimately to be recognized as one of the great books of the world.

The subject of probability is over two hundred years old and for the whole period of its existence there has been dispute about its meaning. At one time these arguments mattered little outside academia, but as the use of probability ideas has spread to so many human activities, and as probabilists have produced more and more sophisticated results, so the arguments have increased in practical importance. Nowhere is this more noticeable than in statistics, where the basic practices of the subject are being revised as a result of disputes about the meaning of probability. When a question has proved to be difficult to answer, one possibility may be that the question itself was wrongly posed and, consequently, unanswerable. This is de Finetti's way out of the impasse. Probability does not exist.

Does not exist, that is, outside of a person: does not exist, objectively. Probability is a description of your (the reader of these words) uncertainty about the world. So this book is about uncertainty, about a feature of life that is so essential to life that we cannot imagine life without it. This book is about life: about a way of thinking that embraces all human activities.

So, in a sense, this book is for everyone; but necessarily it will be of immediate appeal to restricted classes of readers.

Philosophers have recently increased their interest in probability and will therefore appreciate the challenging ideas that the author puts forward. For example, those of the relationships between possibility and tautology. They will notice the continual concern with reality, with the use of the ideas in practical situations. This is a philosophy intended to be operational and to express the individual's appreciation of the external world.

Psychologists are much concerned with the manner of this appreciation, and experiments have been performed which show that individuals do not reason about uncertainty in the way described in these volumes. The experiments provide a descriptive view of man's attitudes: de Finetti's approach is normative. To spend too much time on description is unwise when a normative approach exists, for it is like asking people's opinion of $2+2$, obtaining an average of $4 \cdot 31$ and announcing this to be the sum. It would be better to teach them arithmetic. I hope that this book will divert psychologists' attentions away from descriptions to the important problem, ably discussed in this book, of how to teach people to assess probabilities.

Mathematicians will find much of interest. (Let me hasten to add that some people may approach the book with fear because of the amount of mathematics it contains. They need not worry. Much of the material is accessible with no mathematical skill: yet more needs only a sympathetic appreciation of notation. Even the more mathematical passages use mathematics in a sparse and yet highly efficient way. Mathematics is always the servant – never the master (see Section 1.9.1).) Nevertheless, the mathematician will appreciate the power and elegance of the notation and, in particular, the discussion of finite additivity. He will be challenged by the observation that 'mathematics is an instrument which should conform itself strictly to the exigencies of the field in which it is to be applied'. He will enjoy the new light shed on the calculus of probabilities.

Physicists have long used probabilistic notions in their understanding of the world, especially at the basic, elementary-particle level. Here we have a serious attempt to connect their use of uncertainty with the idea as used outside physics.

Statisticians are the group I can speak about with greatest confidence. They have tended to adopt a view of probability which is based on frequency considerations and is too narrow for many applications. They have therefore been compelled to introduce artificial ideas, like confidence intervals, to describe the uncertainties they need to use. The so-called Bayesian approach has recently made some significant impression, but de Finetti's ideas go further still in replacing frequency concepts entirely – using his notion of exchangeability – and presenting an integrated view of statistics based on a single concept of uncertainty. A consequence of this is that the range of possible applications of statistics is enormously widened so that we can deal with phenomena other than those of a repeatable nature.

There are many other groups of people one would like to see reading these volumes. Operational research workers are continually trying to express ideas to management that involve uncertainty: they should do it using the concepts contained therein. One would like (is it a vain hope?) to see politicians with a sensible approach to uncertainty – what a blessing it would be if they could appreciate the difference between prediction and prevision (p. 60).

The book should therefore be of interest to many people. As the author says (p. 12) 'it is … an attempt to view, in a unified fashion, a group of topics which are in general considered separately, each by specialists in a single field, paying little or no attention to what is being done in other fields.'

The book is not a text on probability in the ordinary sense and would probably not be useful as a basis for a course of lectures. It would, however, be suitable for a graduate seminar wherein sections of it were discussed and analysed. Which sections were used would depend on the type of graduates, but with the continuing emphasis on unity, it would be valuable in bringing different disciplines together. No university should ignore the book.

It would be presumptuous of *me* to say how *you* should read the two volumes but a few words may help your appreciation. Firstly, do not approach it with preconceived ideas about probability. I address this remark particularly to statisticians, who can so easily interpret a formula or a phrase in a way that they have been used to, when de Finetti means something different. Let the author speak for himself. Secondly, the book does not yield to a superficial reading. The author has words of wisdom to say about many things and the wisdom often only appears after reflection. Rather, dip into parts of the book and read those carefully. Hopefully you will be stimulated to read the whole.

Thirdly, the style is refreshing – the translators have cleverly used the phrase 'a whimsical fashion' (Section 1.3.3) – so that every now and again delightful ideas spring to view; the idea that we shall all be Bayesian by 2020, or how-to play the football pools. But, as I said, this is a book about life.

November 1973

University College London,
D.V. Lindley

Preface

I became a postgraduate student of statistics at University College London in 1968, soon after Dennis Lindley had moved there to become the head of the department. He was, at that time, one of the very few academic statisticians committed to the so-called Bayesian approach to the subject. While I was a postgraduate, Lindley several times mentioned to me that his American colleague and fellow Bayesian, L.J. Savage, had encouraged him, and indeed anyone interested in the subjectivist approach to Bayesian statistics, to read the works of the Italian probabilist, actuary and philosopher, Bruno de Finetti.

But there was a problem for most of us at that time. Very little of his work had been translated into English and his 1970 magnum opus, the two-volume *Teoria Delle Probabilitá*, was only available in Italian. The thought of struggling through several hundred pages of dense and difficult writing with the aid of a dictionary was simply too daunting.

In 1971, I left University College London to take up an academic post at the Mathematics Institute in the University of Oxford. Early in 1972, an Italian group theorist called Antonio Machí came to spend a year at the Institute. We became friends and at some stage I mentioned my interest in de Finetti and the frustrations of trying to get to grips with the *Teoria Delle Probabilita*. Antonio immediately suggested that we work together on translating the two-volume work into English. Two years later, after many exchanges between Oxford and Rome, the first Wiley English edition appeared, with a Foreword by Dennis Lindley, with whom I subsequently gave a series of lectures in London to draw the attention of the wider statistics community to the importance of de Finetti's ideas.

There was growing interest in Bayesian ideas throughout the 1970s, but it was still very much a minority view among academic statisticians. The first attempt by some of us to organize a specifically Bayesian international conference in 1978, the first of what were to become the four-yearly Valencia Conferences, attracted around eighty participants. However, by the time we reached the ninth such meeting in 2011, the attendance had grown tenfold and Bayesian thinking had become a significant and influential feature of the statistical landscape.

De Finetti predicts in these volumes that we shall all be Bayesians by 2020. There is still some way to go, but if it proves to be so it will be due in no small measure to the influence of these wonderful volumes.

Adrian Smith

1

Introduction

1.1 Why a New Book on Probability?

There exist numerous treatments of this topic, many of which are very good, and others continue to appear. To add one more would certainly be a presumptuous undertaking if I thought in terms of doing something better, and a useless undertaking if I were to content myself with producing something similar to the 'standard' type. Instead, the purpose is a different one: it is that already essentially contained in the dedication to Beniamino Segre

> [who about twenty years ago pressed me to write it as a necessary document for clarifying one point of view in its entirety.]

Segre was with me at the International Congress of the Philosophy of Science (Paris 1949), and it was on the occasion of the discussions developed there on the theme of probability that he expressed to me, in persuasive and peremptory terms, a truth, perhaps obvious, but which only since appeared to me as an obligation, difficult but unavoidable.

> 'Only a complete treatment, inspired by a well-defined point of view and collecting together the different objections and innovations, showing how the whole theory results in coherence in all of its parts, can turn out to be convincing. Only in this way is it possible to avoid the criticisms to which fragmentary expositions easily give rise since, to a person who in looking for a completed theory interprets them within the framework of a different point of view, they can seem to lead unavoidably to contradictions.'

These are Segre's words, or, at least, the gist of them.

It follows that the requirements of the present treatment are twofold: first of all to clarify, exhaustively, the conceptual premises, and then to give an essentially complete exposition of the calculus of probability and its applications in order to establish the adequacy of the interpretations deriving from those premises. In saying 'essentially' complete, I mean that what matters is to develop each topic just as far as is necessary to avoid conceptual misunderstandings. From then on, the reader could follow any other

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

book without finding great difficulty in making those modifications that are needed in order to translate it, if such be desired, according to the point of view that will be taken here. Apart from these conceptual exigencies, each topic will also be developed, in terms of the content, to an extent sufficient for the treatment to turn out to be adequate for the needs of the average reader.

1.2 What are the Mathematical Differences?

1.2.1. If I thought I were writing for readers absolutely innocent of probabilistic–statistical concepts, I could present, with no difficulty, the theory of probability in the way I judge to be meaningful. In such a case, it would not even have been necessary to say that the treatment contains something new and, except possibly under the heading of information, that different points of view exist. The actual situation is very different, however, and we cannot expect any sudden change.

> My estimation is that another fifty years will be needed to overcome the present situation, but perhaps even this is too optimistic. It is based on the consideration that about thirty years were required for ideas born in Europe (Ramsey, 1926; de Finetti, 1931) to begin to take root in America (even though B.O. Koopman (1940) had come to them in a similar form). Supposing that the same amount of time might be required for them to establish themselves there, and then the same amount of time to return, we arrive at the year 2020.

It would obviously be impossible and absurd to discuss in advance concepts and, even worse, differences between concepts to whose clarification we will be devoting all of what follows; however, much less might be useful (and, anyway, will have to suffice for the time being). It will be sufficient to make certain summary remarks that are intended to exemplify, explain and anticipate for the reader certain differences in attitude that could disorientate him, and leave him undecided between continuing without understanding or, on the other hand, stopping reading altogether. It will be necessary to show that the 'wherefore' exists and to give at least an idea of the 'wherefore', and of the 'wherefores', even without anticipating the 'wherefore' of every single case (which can only be seen and gone into in depth at the appropriate time and place).

1.2.2. From a mathematical point of view, it will certainly seem to the reader that either by desire or through ineptitude I complicate simple things; introducing captious objections concerning aspects that modern developments in mathematical analysis have definitively dealt with. Why do I myself not also conform to the introduction of such developments into the calculus of probability? Is it a question of incomprehension? Of misoneism? Of affectation in preferring to use the tools of the craftsman in an era of automation which allows mass production even of brains – both electronic and human?

The 'wherefore', as I see it, is a different one. To me, mathematics is an instrument that should conform itself strictly to the exigencies of the field in which it is to be applied. One cannot impose, for their own convenience, axioms not required for essential reasons, or actually in conflict with them.

I do not think that it is appropriate to speak of 'incomprehension'. I have followed through, and appreciated, the reasons *pro* (which are the ones usually put forward), but I found the reasons *contra* (which are usually neglected) more valid, and even preclusive.

I do not think that one can talk of misoneism. I am, in fact, very much in favour of innovation and against any form of conservatism (but only after due consideration, and not by submission to the tyrannical caprice of fashion). Fashion has its use in that it continuously throws up novelties, guarding against fossilization; in view of such a function, it is wise to tolerate with goodwill even those things we do not like. It is not wise, however, to submit to passively adapting our own taste, or accepting its validity beyond the limits that correspond to our own dutiful, critical examination.

I do not think that one can talk of 'affectation' either. If anything, the type of 'affectation' that is congenial to my taste would consist of making everything simple, intuitive and informal. Thus, when I raise 'subtle' questions, it means that, in my opinion, one simply cannot avoid doing so.

1.2.3. The 'wherefore' of the choice of mathematical apparatus, which the reader might find irksome, resides, therefore, in the 'wherefores' related to the specific meaning of probability, and of the theory that makes it an object of study. Such 'wherefores' depend, in part, on the adoption of this or that particular point of view with regard to the concept and meaning of probability, and to the basis from which derives the possibility of reasoning about it, and of translating such reasoning into calculations. Many of the 'wherefores' seem to me, however, also to be valid for all, or many, of the different concepts (perhaps with different force and different explanations). In any case, the critical analysis is more specifically hinged on the conception that we follow here, and which will appear more and more clear (and, hopefully, natural) as the reader proceeds to the end – provided he or she has the patience to do so.

1.3 What are the Conceptual Differences?

1.3.1. Meanwhile, for those who are not aware of it, it is necessary to mention that in the conception we follow and sustain here only *subjective* probabilities exist – that is, the *degree of belief* in the occurrence of an event attributed by a given person at a given instant and with a given set of information. This is in contrast to other conceptions that limit themselves to special types of cases in which they attribute meaning to 'objective probabilities' (for instance, cases of symmetry as for dice etc., 'statistical' cases of 'repeatable' events, etc.). This said, it is necessary to add at once that we have no interest, at least for now, either in a discussion, or in taking up a position, about the 'philosophical' aspects of the dispute; in fact, it would be premature and prejudicial because it would entangle the examination of each concrete point in a web of metaphysical misunderstandings.

Instead, we are interested, on the contrary, in clearly understanding what one means *according to one's own conception and in one's own language*, and learning to enter into this conception and language in its motivations and implications (even if provisionally, in order to be able to make pertinent criticism later on). This is, it seems to me, an inviolable methodological need.

1.3.2. There is nothing more disappointing than to hear repeated, presented as 'criticisms', clichés so superficial that it is not possible to infer whether the speaker has even read the arguments developed to confute them and clear them up, or has read them without understanding anything, or else has understood them back to front. The fault could be that of obscure presentation, but a somewhat more meaningful reaction would be required in order to be able to specify accurately, and to correct, those points which lend themselves to misunderstanding.

The fault may be the incompleteness of the preceding, more or less fragmentary, expositions, which, although probably more than complete if taken altogether, are difficult to locate and hold in view simultaneously. If so, the present work should obviate the inconvenience: unfortunately, the fact that it is published is not sufficient; the result depends on the fact that it is read with enough care to enable the reader to make pertinent criticisms.

I would like to add that I understand very well the difficulties that those who have been brought up on the objectivistic conceptions meet in escaping from them. I understand it because I myself was perplexed for quite a while some time ago (even though I was free from the worst impediment, never having had occasion to submit to a ready-made and presented point of view, but only coming across a number of them while studying various books and works on my own behalf). It was only after having analysed and mulled over the objectivistic conceptions in all possible ways that I arrived, instead, at the firm conviction that they were all irredeemably illusory. It was only after having gone over the finer details and developed, to an extent, the subjectivistic conception, assuring myself that it accounted (in fact, in a perfect and more natural way) for everything that is usually accredited, overhastily, to the fruit of the objectivistic conception, it was only after this difficult and deep work, that I convinced myself, and everything became clear to me. It is certainly possible that these conclusions are wrong; in any case they are undoubtedly open to discussion, and I would appreciate it if they were discussed.

However, a dialogue between the deaf is not a discussion. I think that I am doing my best to understand the arguments of others and to answer them with care (and even with patience when it is a question of repeating things over and over again to refute trivial misunderstandings). It is seldom that I have the pleasure of forming the impression that other people make a similar effort; but, as the Gospel says, 'And why beholdest thou the mote that is in thy brother's eye, but considerest not the beam that is in thine own eye?': if this has happened to me, or is happening to me, I would appreciate it if someone would enlighten me.

1.3.3. One more word (hopefully unnecessary for those who know me): I find it much more enlightening, persuasive, and in the end more essentially serious, to reason by means of paradoxes; to reduce a thesis to absurdity; to make use of images, even lighthearted ones provided they are relevant, rather than to be limited to lifeless manipulations in technical terms, or to heavy and indigestible technical language. It is for this reason that I very much favour the use of colourful and vivid forms of expression, which, hopefully, may turn out to be effective and a little entertaining, making concrete, in a whimsical fashion, those things that would appear dull, boring or insipid and, therefore, inevitably badly understood, if formulated in an abstract way, stiffly or with affected gravity. It is for this reason that I write in such a fashion, and desire to do so; not because of ill-will or lack of respect for other people, or their opinions (even when I judge them wrong). If somebody finds this or that sentence a little too sharp, I beg him to believe in the total absence of intention and animosity, and to accept my apologies as of now.

1.4 Preliminary Clarifications

1.4.1. For the purpose of understanding, the important thing is not the difference in philosophical position on the subject of probability between 'objective' and 'subjective', but rather the resulting reversals of the rôles and meanings of many concepts, and, above all, of what is 'rigorous', both logically and mathematically. It might seem paradoxical but the fact is that the subjectivistic conception distinguishes itself precisely by a more rigorous respect for that which is really objective, and which it calls, therefore, 'objective'.[1] There are cases in which, in order to define a notion, in formulating the problem, or in justifying the reasoning, there exists a choice between an unexceptionable, subjectivistic interpretation and a would-be objectivistic interpretation. The former is made in terms of the opinions or attitudes of a given person; the latter derives from a confused transposition from this opinion to the undefinable complex of objective circumstances that might have contributed to its determination: in such cases there is nothing to do but choose the first alternative. The subjective opinion, as something known by the individual under consideration, is, at least in this sense, something objective and can be a reasonable object of a rigorous study. It is certainly not a sign of greater realism, of greater respect for objectivity, to substitute for it a metaphysical chimera, even if with the laudable intention of calling it 'objective' in order to be able to then claim to be concerned only with objective things.

There might be an objection that we are in a vicious circle, or engaged in a vacuous discussion, since we have not specified what is to be understood by 'objective'. This objection is readily met, however: statements have *objective* meaning if one can say, on the basis of a well-determined observation (which is at least conceptually possible), whether they are either TRUE or FALSE. Within a greater or lesser range of this delimitation a large margin of variation can be tolerated, with one condition – do not *cheat*. To *cheat* means to leave in the statement sufficient confusion and vagueness to allow ambiguity, second-thoughts and equivocations in the ascertainment of its being TRUE or FALSE. This, instead, must always appear simple, neat and definitive.

1.4.2. Statements of this nature, that is the only 'statements' in the true sense of the word, are the object of the *logic of certainty*, that is ordinary logic, which could also be in the form of mathematical logic, or of mathematics. They are also the objects *to which* judgements of probability apply (as long as one does not know whether they are true or false) and are called either *propositions*, if one is thinking more in terms of the expressions in which they are formulated, or *events*, if one is thinking more in terms of the situations and circumstances to which their being true or false corresponds.

On the basis of the considerations now developed, one can better understand the statement made previously, according to which the fundamental difference between the subjectivistic conception and the objectivistic ones is not philosophical but *methodological*. It seems to me that no-one could refute the methodological rigour of the subjectivistic conception: not even an objectivist. He himself, in fact, would have unlimited need of it in trying to expose, in a sensible way, the reasons that would lead him to consider 'philosophically correct' this one, or that one, among the infinitely many possible opinions about the evaluations of probability. To argue against this can only mean,

1 This fact has often been underlined by L.J. Savage (see Kyburg and Smokler (1964), p. 178, and elsewhere).

even though without realizing it, perpetuating profitless discussions and playing on the ambiguities that are deeply rooted in the uncertainty.

At this stage, a few simple examples might give some preliminary clarification of the meaning and compass of the claimed 'methodological rigour' – under the condition, however, that one takes into account the necessarily summary character of these preliminary observations. It is necessary to pay attention to this latter remark to avoid both the acceptance of such observations as exhaustive and the criticism of them that results from assuming that they claim to be exhaustive: one should realize, with good reason, that they are by no means such.

1.5 Some Implications to Note

1.5.1. We proceed to give some examples: to save space, let us denote by 'O' statements often made by *objectivists*, and by 'S' those with which a *subjectivist* (or, anyway, *this author*) would reply.

O: Two events of the same type in identical conditions for all the relevant circumstances are 'identical' and, therefore, necessarily have the same probability.[2]

S: Two distinct events are always different, by virtue of an infinite number of circumstances (otherwise how would it be possible to distinguish them?!). They are equally probable (for an individual) if – and so far as – he judges them as such (possibly by judging the differences to be irrelevant in the sense that they do not influence his judgement). An even more fundamental objection should be added: the judgement about the probability of an event depends not only on the event (or on the person) but also on the state of information. This is occasionally recalled, but more often forgotten, by many objectivists.

O: Two events are (stochastically) independent[3] if the occurrence of one does not influence the probability of the other.

S: I would say instead: by definition, two events are such (for an individual) if the knowledge of the outcome of one does not make him change the evaluation of probability for the other.

O: Let us suppose *by hypothesis* that these events are equally probable, for example with probability $p = \frac{1}{2}$, and independent, and so on.

S: It is meaningless to consider as an 'hypothesis' something that is not an objective statement. A statement about probability (the one given in the example or any other one whatsoever) either *is* the evaluation of probabilities (those of the speaker or of someone else), in which case there is nothing to do but simply register the fact, or it is nothing.

O: These events are independent and all have the same probability which is, however, 'unknown'.

---

2 The objectivists often use the word event in a generic sense also, using '*trials*' (or '*repetitions*') *of the same* '*event*' to mean *single events*, '*identical*' *or* '*similar*'. From time to time we will say '*trials*' (or '*repetitions*') *of a phenomenon*, always meaning by event a single event. It is not simply a question of terminology, however: we use 'phenomenon' because we do not give this word any technical meaning; by saying 'trials of a phenomenon' one may allude to some exterior analogy but one does not mean to assume anything that would imply either equal probability, or independence, or anything else of probabilistic relevance.

3 Among events, random quantities, or random entities in general, it is possible to have various relations termed 'independence' (linear, logical, stochastic); it is better to be specific if there is any risk of ambiguity.

S: This formulation is a nonsense in the same sense as the preceding one but to a greater extent. By interpreting the underlying intention (which, as an intention, is reasonable) one can translate it (see Chapter 11) into a completely different formulation, 'exchangeability', in which we do *not* have independence, the probabilities are *known*, and vary, precisely, in depending only on the number of successes and failures of which one has information.

One might continue in this fashion, and it could be said that almost the whole of what follows will be, more or less implicitly, a continuation of this same discussion. Rather, let us see, by gathering together the common factors, the essential element in all these contrapositions.

1.5.2. For the subjectivist everything is clear and rigorous when he is expressing something about somebody's evaluation of probabilities; an evaluation which is, simply, what it is. For that somebody, it will have motivations that we might, or might not, know; share, or not share; judge[4] more or less reasonable, and that might be more or less 'close' to those of a few, or many, or all people. All this can be interesting, but it does not alter anything. To express this in a better way: all these things matter in so far as they determined that unique thing that matters, and that is the evaluation of probability to which, in the end, they have given rise.

From the theoretical, mathematical point of view, even the fact that the evaluation of probability expresses somebody's opinion is then irrelevant. It is purely a question of studying it and saying whether it is coherent or not; that is whether it is free of, or affected by, intrinsic contradictions. In the same way, in the logic of certainty one ascertains the correctness of the deductions but not the accuracy of the factual data assumed as premises.

1.5.3. Instead, the objectivist would like to ignore the evaluations, actual or hypothetical, and go back to the circumstances that might serve as a basis for motivations which would lead to evaluations. Not being able to invent methods of synthesis comparable in power and insight to those of the human intuition, nor to construct miraculous robots capable of such, he contents himself, willingly, with simplistic schematizations of very simple cases based on neglecting all knowledge except a unique element which lends itself to utilization in the crudest way.

A further consequence is the following. The subjectivist, who knows how much caution is necessary in order to remain within the bounds of realism, will exercise great care in not going far beyond the consideration of cases immediately at hand and directly interesting. The objectivist, who substitutes the abstraction of schematized models for the changing and transient reality, cannot resist the opposite temptation. Instead of engaging himself, even though in a probabilistic sense (the only one which is valid), in saying something about the specific case of interest, he prefers to 'race on ahead', occupying himself with the asymptotic problems of a large number of cases, or even playing around with illusory problems, contemplating infinite cases where he can try, without any risk, to pass off his results as 'certain predictions'.[5]

4 With a judgment which is 'subjective squared': our subjective judgment regarding the subjective judgment of others.

5 Concerning the different senses in which we use the terms 'prevision' and 'prediction', see Chapter 3 (at the beginning and then in various places, in particular 3.7.3).

1.6 Implications for the Mathematical Formulation

1.6.1. From these conceptual contrapositions there follows, amongst other things, an analogous contraposition in the way in which the mathematical formulation is conceived. The subjectivistic way is the one that it seems appropriate to call 'natural': it is possible to evaluate the probability over any set of events whatsoever; those for which it serves a purpose, or is of interest, to evaluate it; there is nothing further to be said. The objectivistic way (and also the way most congenial to contemporary mathematicians, independently of the conception adopted regarding probability) consists in requiring, as an obligatory starting point, a mathematical structure much more formidable, complete and complicated than necessary (and than it is, in general, reasonable to regard as conceivable).

1.6.2. Concerning a known evaluation of probability, over any set of events whatsoever, and interpretable as the opinion of an individual, real or hypothetical, we can only judge whether, or not, it is *coherent*.[6] If it is not, the evaluator, when made aware of it, should modify it in order to make it coherent. In the same way, if someone claimed to have measured the sides and area of a rectangle and found 3 m, 5 m and 12 m^2, we, even without being entitled, or having the inclination, to enter into the merits of the question, or to discuss the individual measurements, would draw his attention to the fact that at least one of them is wrong, since it is not true that $3 \times 5 = 12$.

Such a condition of coherence should, therefore, be *the weakest one* if we want it to be the strongest in terms of absolute validity. In fact, *it must only exclude the absolutely inadmissible evaluations*; that is those that one cannot help but judge contradictory (in a sense that we shall see later).

Such a condition, as we shall see, reduces to *finite additivity* (and *non-negativity*). It is not admissible to make it more restrictive (unless it turns out to be necessary if we discover the preceding statement to be wrong); it would make us exclude, erroneously, admissible evaluations.

1.6.3. What the objectivistic, or the purely formalistic, conceptions generally postulate is, instead, that countable additivity holds (as for Borel or Lebesgue measure), and that the field over which the probability is defined be the whole of a Boolean algebra. From the subjectivistic point of view this is both too much and too little: according to what serves the purpose and is of interest, one could limit oneself to much less, or even go further. One could attribute probabilities, finitely but not countably additive, to all, and only, those events that it is convenient to admit into the formulation of a problem and into the arguments required for its solution. One might also go from one extreme to the other: referring to the analogy of events and probability with sets and measure, it might, at times, be convenient to limit oneself to thinking of a measure as defined on certain simple sets (like the intervals), or even on certain sets but not their intersections (for instance, for 'vertical' and 'horizontal' 'stripes' in the (x,y)-plane ($x' \leq x < x''$, $y' \leq y < y''$) but not on the rectangles); and, at other times, to think of it instead as extended to all the sets that the above-mentioned convention would exclude (like the 'non-Lebesgue-measureable sets').

---

6 See Chapter 3.

1.6.4. In a more general sense, it seems that many of the current conceptions consider as a success the introduction of mathematical methods so powerful, or of tricks of formulation so slick, that they permit the derivation of a uniquely determined answer to a problem even when, due to the insufficiency of the data, it is indeterminate. A capable geometer in order to conform to this aspiration would have to invent a formula for calculating the area of a triangle given two sides.

> Attempts of this kind are to be found in abundance, mainly in the field of statistical induction (see some remarks further on in this Introduction, 1.7.6).
>
> In the present case, the defect is somewhat hidden and consists in the following distinction between the two cases of *measure* and of *probability*.
>
> To extend a *mathematical* notion (measure) from one field (Jordan–Peano) to another (Borel–Lebesgue) is a question of convention. If, however, a notion (like probability) *already has a meaning* (for each event, at least potentially, even if not already evaluated), one cannot give it a value by conventional extension of the probabilities already evaluated *except for the case in which it turns out to be the unique one compatible with them by virtue of the sole conditions of coherence* (conditions pertaining to the meaning of probability, not to motives of a mathematical nature). The same would happen if it were a question of a physical quantity like mass. If one thought of being able to give meaning to the notion of 'mass belonging to any set of points of a body' (for instance those with rational coordinates), in the sense that it were, at least conceptually, possible to isolate such a mass and weigh it, then it would be legitimate, when referring to it, to talk about everything that can be deduced about it by mathematical properties that translate necessary physical properties, and only such things. To say something more (and in particular to give it a unique value when such properties leave the value indeterminate between certain limits), by means of the introduction of arbitrary mathematical conventions, would be unjustified, and therefore inadmissible.

1.7 An Outline of the 'Introductory Treatment'

1.7.1. The reader must feel as though he has been plunged alternately into baths of hot and cold water: in Section 1.5 he encountered the contraposed examples of the conceptual formulation, presented either as meaningful or as meaningless; in Section 1.6 the mathematical formulations, presented either as suitable or as academic. Following this, a simple and ordered presentation of the topics that will follow may provide a suitable relaxation, and might even induce a return to the preceding 'baths' in order, with a greater knowledge of the motives, to soak up some further meaning.

1.7.2. In Chapter 2 we will *not* talk of probability. Since we wish to make absolutely clear the distinction between the subjective character of the notion of probability and the objective character of the elements (events, or any random entities whatsoever) to which it refers, we will first treat only these entities. In other words, we will deal with the preliminary logic of certainty where there exist only:

- TRUE and FALSE as final answers;
- CERTAIN and IMPOSSIBLE and POSSIBLE as alternatives, with respect to the present knowledge of each individual.

In this way, the range of uncertainty, that is of what is not known, will emerge in outline. This is the framework into which the (subjective) notion of probability will be introduced as an indispensable tool for our orientation and decision making.

The random events, random quantities and any other random entities, will already be defined, however, before we enter the domain of probability, and they will simply be events, quantities, entities, well-defined but with no particular features except the fact of not being known by a certain individual. For any individual who does not know the value of a quantity X, there will be, instead of a unique *certain* value, two, or several, or infinitely many, *possible* values of X. They depend on his degree of ignorance and are, therefore, relative to his state of information; nevertheless, they are objective because they do not depend on his opinions but only on these objective circumstances.

1.7.3. Up until now the consideration of uncertainty has been limited to the negative aspect of *nonknowledge*. In Chapter 3 we will see how the need arises, as natural and appropriate, to integrate this aspect with the positive aspect (albeit weak and temporary while awaiting the information that would give it certainty) given by the evaluation of probabilities. To any event in which we have an interest, we are accustomed to attributing, perhaps vaguely and unconsciously, a probability: if we are sufficiently interested we may try to evaluate it with some care. This implies introspection in depth by weighing each element of judgment and controlling the coherence by means of other evaluations made with equal accuracy. In this way, each event can be assigned a probability, and each random quantity or entity a distribution of probability, as an expression of the attitude of the individual under consideration.

Let us note at once a few of the points that arise.

Others, in speaking of a random quantity, assume a probability distribution as already attached to it. To adopt a different concept is not only a consequence of the subjectivistic formulation, according to which the distribution can vary from person to person, but also of the unavoidable fact that the distribution varies with the information (a fact which, in any case, makes the usual terminology inappropriate).

Another thing that might usefully be mentioned now is that the conditions of coherence will turn out to be particularly simplified and clarified by means of a simple device for simultaneously handling events and random quantities (or entities of any linear space whatever). Putting the logical values 'True' and 'False' equal to the numbers '1' and '0', an event is a random quantity that can assume these two values: the function $\mathbf{P}(X)$, which for X = event gives its probability, is, for arbitrary X, the 'prevision' of X (i.e. in the usual terminology, the mathematical expectation).

The use of this *arithmetic* interpretation of the events, preferable to, but not excluding, the set-theoretic interpretation, has its utility and motivation, as will be seen. The essential fact is that the *linearity* of the arithmetic interpretation plays a fundamental rôle (which is, in general, kept in the background), whereas the structure of the Boolean algebra enters rather indirectly.

1.7.4. After having extended these considerations, in Chapter 4, to the case of conditional probabilities and previsions (encountering the notions of stochastic independence and correlation), we will, in Chapter 5, dwell upon the evaluation of probabilities. The notions previously established will allow us not only to apply the instruments for this evaluation, but also to relate them to the usual criteria, inspired by partial, objectivistic

'definitions'. We will see that the subjectivistic formulation, far from making the valid elements in the ideas underlying these criteria redundant, allows the best and most complete use of them, checking and adapting, case by case, the importance of each of them. In contrast to the usual, and rather crude, procedure, which consists of the mechanical and one-sided application of this or that criterion, the proposed formulation allows one to behave in conformity with what the miraculous robot, evoked in Section 1.5.3, would do.

1.7.5. Chapters 6–10 extend to give a panoramic vision of the field of problems with which the calculus of probabilities is concerned. Of course, it is a question of compromising between the desire to present a relatively complete overall view and the desire to concentrate attention on a small number of concepts, problems and methods, whose rôle is fundamental both in the first group of ideas, to be given straightaway, and, even more, in further developments, which, here, we can at most give a glimpse of.

Also in these chapters, which in themselves are more concerned with content than critical appraisal, there are aspects and, here and there, observations and digressions that are relevant from the conceptual angle. It would be inappropriate to make detailed mention of them but, as examples, we could quote the more careful analysis of what the knowledge of the distribution function says, or does not say (also in connection with the 'possible' values), and of the meaning of 'stochastic independence' (between random quantities), expressed by means of the distribution function.

1.7.6. The last two chapters, 11 and 12, deal briefly with the problems of induction (or inference) and their applications, which constitute mathematical statistics. Here we encounter anew the conceptual questions connected with the subjective conception, which, of course, bases all inference on the Bayesian procedure (from Thomas Bayes,[7] 1763). In this way, the theory and the applications come to have a unified and coherent foundation: it is simply a question of starting from the evaluation of the initial probabilities (i.e. before acquiring new information – by observation, experiment, or whatever) and then bringing them up to date on the basis of this new information, thus obtaining the final probabilities (i.e. those on which to base oneself after acquiring such information).

The objectivistic theories, in seeking to eschew the evaluation and use of 'initial probabilities', lack an indispensable element for proceeding in a sensible way and appeal to a variety of empirical methods, often invented *ad hoc* for particular cases. We shall use the term '*Adhockeries*', following Good[8] (1965) who coined this apt expression, for the methods, criteria and procedures that, instead of following the path of the logical formulation, try to answer particular problems by means of particular tricks (which are sometimes rather contrived).

---

7 One must be careful not to confuse Bayes' *theorem* (which is a simple corollary of the theorem of compound probabilities) with Bayes' *postulate* (which assumes the uniform distribution as a representation of 'knowing nothing'). Criticisms of the latter, often mistakenly directed against the former, are not therefore valid as criticisms of the position adopted here.

8 Good's position is less radical than I supposed when I interpreted 'Adhockery' as having a derogatory connotation. I gathered this from his talk at the Salzburg Colloquium, and commented to this effect in an Addendum to the paper I delivered there; *Synthese* **20** (1969), 2–16: 'According to it, "adhockeries" ought not to be rejected outright; their use may sometimes be an acceptable substitute for a more systematic approach. I can agree with this only if – and in so far as – such a method is justifiable as an approximate version of the correct (i.e. Bayesian) approach. (Then it is no longer a mere "adhockery".)'

1.8 A Few Words about the 'Critical' Appendix

1.8.1. Many of the conceptual questions are, unfortunately, inexhaustible if one wishes to examine them thoroughly; and the worst thing is that, often, they are also rather boring unless one has a special interest in them.

A work that is intended to clarify a particular conceptual point of view cannot do without this kind of analysis in depth, but it certainly seems appropriate to avoid weighing down the text more than is necessary to meet the needs of an ordinary reader who desires to arrive at an overall view. For this reason, the most systematic and detailed critical considerations have been postponed to an Appendix. This is intended as a reassurance that there is no obligation to read it in order to understand what follows, nor to make the conclusions meaningful. This does not mean, however, that it is a question of abstruse and sophisticated matters being set aside for a few specialists and not to be read by others. It is a question of further consideration of different points that might appear interesting and difficult, to a greater or lesser extent, but which might always improve, in a meaningful and useful, though not indispensable, way, the awareness of certain questions and difficulties, and of the motives which inspire different attitudes towards them.

1.8.2. In any case, one should point out that it is a question of an attempt to view, in a unified fashion, a group of topics that are in general considered separately, each by specialists in a single field, paying little or no attention to what is being done in other fields. Notwithstanding the many gaps or uncertainties, and the many imperfections (and maybe precisely also for the attention it may attract to them), I think that such an attempt should turn out to be useful.

Among other things, we have tried to insert into the framework of the difficulties associated with the 'verifiability' of events in general, the question of 'complementarity' that arose in quantum physics. The answer is the one already indicated, in a summary fashion, elsewhere (de Finetti, 1959), and coinciding with that of B.O. Koopman (1957), but the analysis has been pursued in depth and related to the points of view of other authors as far as possible (given the margin of uncertainty in the interpretation of the thought of those consulted, and the impossibility of spending more time on this topic in attempting to become familiar with others).

1.8.3. Various other questions that are discussed extensively in the Appendix, are currently objects of discussion in various places: for instance, the relationships between possibility and tautology seem to be attracting the attention of philosophers (the intervention of Hacking at a recent meeting, Chicago 1967); while the critical questions about the mathematical axioms of the calculus of probability (in the sense, to be understood, of making it a theory strictly identical to measure theory, or with appropriate variations) are always a subject of debate.

Apart from the points of view on separate questions, the Appendix will also have as a main motive the proposal to model the mathematical formulation on the analysis of the actual needs of the substantive interpretation. Moreover, to do so with the greatest respect for 'realism', which the inevitable degree of idealization must purify just a little, but must never overwhelm or distort, neither for analytical convenience, nor for any other reason.

1.9 Other Remarks

1.9.1. It seems appropriate here to draw attention also to some further aspects, all secondary, even if only to underline the importance that attaches, in my opinion, to 'secondary' things.

One characteristic of the calculus of probability is that mathematical results are often automatically obtained because their probabilistic interpretations are obvious. In all these cases I think it is much more effective and instructive to consider as their proofs these latter expressive interpretations, and as formal verifications their translation into technical details (to be omitted, or left to the reader). This seems to me to be the best way of realizing the ideal expressed in the maxim that Chisini[9] often repeated: '*mathematics is the art which teaches one how not to make calculations*'.

It is incredible how many things are regularly presented in a heavy and obscure fashion, arriving at the result through a labyrinth of calculations that make one lose sight of the meaning, whereas simple, synthetic considerations would be sufficient to reveal that, for those not wishing to behave as if handcuffed or blindfolded, results and meaning are at hand, staring one in the face.

On numerous occasions one sees very long calculations made in order to prove results that are either wrong or obvious. The latter case is the more serious, without any extenuating circumstances, since it implies lack of realization that the conclusion was obvious, even after having seen it. On the other hand, failing to get the result due to a casual mistake merits only half a reproach, since the lack of realization only applies before starting the calculations.

Instead, it is often sufficient to remark that two formulae are necessarily identical for the simple reason that they express the same thing in different ways, since they provide the result of the same process starting from different properties which characterize it, or for other similar reasons. Problems that can, more or less 'surprisingly', be reduced to synthetic arguments arise frequently in, amongst other things, questions connected with random processes (ranging from the game of Heads and Tails to cases involving properties of characteristic functions etc.). Often, on the other hand, it is an appropriate geometric representation that clarifies the situation and also suggests, without calculations and without any doubts, the solution in formulae.

1.9.2. In addition, however, there are even more secondary things which have their importance. These I would like to explain with a few examples so that it does not seem that some small innovation, perhaps in notation or terminology, has been introduced just for the sake of changing things, instead of with reluctance, overcome by the realization that this was the only way of getting rid of many useless complications.

The very simple device, from which most of the others derive, is that mentioned already in 1.7.3. We identify an event E with the random quantity, commonly called the 'indicator of E', which takes values 1 or 0 according to whether E is true or false. Not only can one operate arithmetically on the events (the arithmetic sum of many events = the

9 Oscar Chisini, a distinguished and gifted pupil of Federigo Enriques, was Professor at the University of Milan where the author attended his course on Advanced Geometry. Chisini's generalized definition of the concept of mean (see Chapter 2, Section 2.9) came about as a result of his occasionally being concerned with this notion in connection with secondary-school examinations.

number of successes; $E - p$ = the gain from a bet for a person who stakes a sum p in order to receive a sum 1 if E occurs etc.) but one operates with a unique symbol **P** in order to denote both probability and prevision (or 'mathematical expectation'), thus avoiding duplication. The 'theorem' $\mathbf{M}(I_E) = \mathbf{P}(E)$, 'the mathematical expectation of the indicator of an event is equal to the probability of the same event', *is rendered superfluous* (it could only be expressed by $\mathbf{P}(E) = \mathbf{P}(E)$!).

1.9.3. The identification TRUE = 1, FALSE = 0 is also very useful as a simple conventional device for denoting, in a straightforward and synthetic way, many mathematical expressions that usually require additional verbal explanation. Applying the same identifications to formulae expressing conditions, for instance, interpreting '$(0 \leqslant x \leqslant 1)$' as a symbol with value 1 for x between 0 and 1, where the inequality is *true*, and value 0 outside, where it is *false*, one can simply write expressions of the type

$$f(x) = g(x) \qquad \left(0 \leqslant x \leqslant 1\right)$$

(and more complicated forms), which otherwise require verbal explanations, like 'the function $f(x)$ which coincides with $g(x)$ for $0 \leqslant x \leqslant 1$ and is zero elsewhere', or writing in the cumbersome form

$$f(x) = \begin{cases} = 0 & \text{for} \quad x < 0, \\ = g(x) & \text{for} \quad 0 \leqslant x \leqslant 1, \\ = 0 & \text{for} \quad x > 1. \end{cases}$$

It is easy to imagine many cases in which the utility of such a convention is much greater, but I think it is difficult to realize the number and variety of such cases (I am often surprised by new, important applications not previously foreseen).

1.9.4. Other simplifications of this kind, which can sometimes be used in conjunction with the above, result from a parallel (or dual) extension of the Boolean operations to the field of real numbers, coinciding, for the values 0 and 1, with the usual meaning for the events. This natural and meaningful extension will also reveal its utility in many applications[10] (see Chapter 2, Sections 2.5 and 2.11).

1.9.5. A small innovation in notation is that of denoting the three most important types of convergence in the probabilistic field by:

symbol: type of convergence:

$\overset{<}{\rightarrow}$	weak	(in probability)	(in measure)
$\overset{>}{\rightarrow}$	strong	(almost certain)	(almost everywhere)
$\overset{\cdot}{\rightarrow}$	quadratic	(in mean-square)	(in mean (quadratic))

10 The advantages of these two conventions (0 and 1 for True–False, and $\vee$ and $\wedge$ among numbers) are illustrated, somewhat systematically and with concise examples, in a paper in the volume in honour of O. Onicescu (75th birthday): 'Revue roumaine de mathématiques pures et appliquées', Bucharest (1967), **XII**, 9, 1227–1233. An English translation of this appears in B. de Finetti, *Probability, Induction and Statistics*, John Wiley & Sons (1972).

(this could also have value in function theory). The innovation seems to me appropriate not only to avoid abbreviations which differ from language to language but also for greater clarity, avoiding the typographic composition and deciphering of symbols that are either cumbersome or unreadable.

1.9.6. Another device which we will introduce with the intention of simplifying the notation does not have a direct relationship with the calculus of probability. For this reason we were even more hesitant to introduce it, but finally realized that without such a remedy there remained simple and necessary things which could not be expressed in a decently straightforward way.

The most essential is the device of obtaining symbols indicating functions, substituting for the variable (in any expression whatever) a 'place-name' symbol: as such □ would seem suitable; it also suggests something which awaits filling in. The scope is the same as obtained by Peano by means of the notation '$|x$', 'varying x', which, applied for instance to the expression $(x \sin x^2 + \sqrt{(3 - x)})/\log(2 + \cos x)$ gives $f = \{[(x \sin x^2 + \sqrt{(3 - x)})/\log(2 + \cos x)]|x\}$, where f is the symbol of the function such that $f(x)$ gives the expression above, and $f(y)$, $f(ax^2 + b)$, $f(e^z)$, … is the same thing in which at each place where an x is found we substitute y, or $ax^2 + b$, or e^z, or whatever. This notation, however, does not lend itself to many cases where it would be required, and where, instead, the notation which puts the 'place-name' for the variable, which is left at our disposal,[11] is very useful. In the preceding example one would write

$$f = \frac{\square \sin \square^2 + \sqrt{\left(3 - \square\right)}}{\log\left(2 + \cos \square\right)}$$

and to denote $f(x)$, $f(y)$, $f(ax^2 + b)$, $f(e^z)$, it would suffice to write on the right, within parentheses, (), the desired variable.

The greatest utility is perhaps obtained in the simplest cases: for instance, in order to denote by □, $\square^2$, $\square^{-1}$ the identity function, $f(x) = x$, or the quadratic, $f(x) = x^2$, or the reciprocal, $f(x) = 1/x$, when the f must be denoted as the argument in a functional. For example, $F(\square)$, $F(\square^2)$, might indicate the first and second moments of a distribution F (according to the conventions of which we shall speak in Chapter 6), and then for any others, $F(\square^n)$, $F(|\square^n|)$ and so on.

1.9.7. Finally, a secondary device is that of consistently denoting by K any multiplicative constant whatever and, if necessary, indicating its expression immediately afterwards, instead of writing it directly, in extensive form, in the formulae. Otherwise, it often happens that a function, of x say, has a rather complicated appearance and each symbol, even those in small print or in the exponents and so on, must be deciphered with care in order to see where x appears. Often one subsequently realizes that the function is very simple and that the complexity of the expression derives solely from having expressed the constant in extensive form. We may have a normalizing constant which, at times, could even be ignored because it automatically disappears in the sequel, or can be calculated more

11 In the case of many variables (for instance three) one could easily use the same device, putting in their places different 'placenames'; for example, $\square_1$, $\square_2$, $\square_3$, with the understanding that $f(x, y, z)$ or $f(5, -\frac{1}{2}, 0)$ or $f(x + y, -\frac{1}{2}x, 1 - 2y)$ etc., is what one obtains putting the 1st or 2nd or 3rd elements of the triple in the places indicated by the three 'place names' with indices 1, 2, 3.

easily from the final formula. At times, in fact, it will be left as a 'reminder' of the existence of an omitted multiplicative factor, which will always be indicated by K, even if the value might change at each step: the reader should make careful note of this remark.

1.10 Some Remarks on Terminology

1.10.1. It is without doubt unreasonable, and rather annoying, to dwell at length on questions of terminology; on the other hand, a dual purpose glossary would be useful and instructive. In the first place, it could improve on a simple alphabetical index in aiding those who forget a definition, or remember it only vaguely; secondly, it could explain the motivation behind the choice, or sometimes the creation, of certain terms, or the fixing of certain conventions for their use.[12] For those interested, such an explanation would also provide an account of the wherefores of the choices. Such a glossary would, however, be out of place here and, in any case, the unusual terms are few and they will be explained as and when they arise.

1.10.2. More importantly, attention must be drawn to some generic remarks, like paying attention to the nuances of divergences of interpretation, which depend on differences in conception. The main one, that of registering that an *event* is always a single case, has already been underlined (Section 1.5.1); the same remark holds for a *random quantity* (Section 1.7.2), and for every kind of 'random entity'. Two clarifications of terminology are appropriate at this juncture: the first to explain why I do not use the term 'variable'; the second to explain the different uses of the terms 'chance', 'random' and 'stochastic'.

To say 'random (or "chance") variable' might suggest that we are thinking of the 'statistical' interpretation in which one thinks of many 'trials' in which the random quantity can *vary*, assuming different values from trial to trial: this is contrary to our way of understanding the problem. Others might think that, even if it is a question of a unique well-determined value, it is 'variable' for one who does not know it, in the sense that it may assume any one of the values 'possible' for him. This does not appear, however, to be a happy nomenclature, and, even less, does it appear to be necessary. In addition, if one wanted to adopt it, it would be logical to do so always, by saying: random variable numbers, random variable vectors, random variable points, random variable matrices, random variable distributions, random variable functions, …, random variable events, and not saying random vector, random point, random matrix, random distribution, random function, random event, and only in the case of numbers not to call it number any more, but variable.

With regard to the three terms – '*chance*', '*random*', '*stochastic*' – there are no real problems: it is simply the convenience of avoiding indiscriminate usage by supporting the consolidation of a tendency that seems to me already present but not, as far as I know, expressly stated. Specifically, it seems to me preferable to use, systematically:

- '*Random*' for that which is the *object* of the theory of probability (as in the preceding cases); I will, therefore, say random process, not stochastic process.
- '*Stochastic*' for that which is valid 'in the sense of the calculus of probability': for instance, stochastic independence, stochastic convergence, stochastic integral; more

12 A very good example would be that of the *Dictionary* at the end of the 'book' by Bourbaki (1939).

generally, stochastic property, stochastic models, stochastic interpretation, stochastic laws; or also, stochastic matrix, stochastic distribution,[13] and so on.
- '*Chance*' is perhaps better reserved for less technical use: in the familiar sense of 'by chance', 'not for a known or imaginable reason', or (but in this case we should give notice of the fact) in the sense of 'with equal probability' as in 'chance drawings from an urn', 'chance subdivision', and similar examples.

1.10.3. Special mention should be made of what is perhaps the important change in terminology: *prevision* in place of mathematical expectation, or expected value and so on. Firstly, all these other nomenclatures have, taken literally, a rather inappropriate meaning and often, through the word 'expectation', convey something old-fashioned and humorous (particularly in French and Italian, where 'espérance' and 'speranza' primarily mean 'hope'!). In any case, it is inconvenient that the expression of such a fundamental notion, so often repeated, should require two words. Above all, however, there was another reason: to use a term beginning with *P*, since the symbol **P** (from what we have said and recalled) then serves for that unique notion which in general we call *prevision*[14] and, in the case of events, also *probability*.[15]

1.11 The Tyranny of Language

All the devices of notation and terminology and all the clarifications of the interpretations are not sufficient, however, to eliminate the fundamental obstacle to a clear and simple explication, adequate for conceptual needs: they can at most serve as palliatives, or to eliminate blemishes.

That fundamental obstacle is the difficulty of escaping from the tyranny of everyday language, whose viscosity often obliges us to adopt phrases conforming to current usage instead of meditating on more apt, although more difficult, versions. We all continue to say 'the sun rises' and I would not know which phrase to use in order not to seem an anachronistic follower of the Ptolemaic system. Fortunately the suspicion does not even enter one's mind because nobody quibbles about the literal meaning of this phrase.

---

13 The case of *matrices* and *distributions* illustrates the difference well. A random matrix is a matrix whose entries are random quantities; a stochastic matrix (in the theory of Markov chains) is the matrix of 'transition probabilities'; i.e. well-determined quantities that define the random process. A random distribution (well-defined but not known) is that of the population in a future census, according to age, or that of the measures that will be obtained in *n* observations that are to be made; a stochastic distribution would mean distribution of probability (but it is not used, nor would it be useful).

14 *Translators' note*. We have used *prevision* rather than *foresight* (as in Kyburg and Smokier, p. 93) precisely for the reasons given in 1.10.3.

15 SIn almost all languages other than Italian, the letter *E* is unobjectionable, and often a single word is sufficient: Expectation (English), Erwartung (German), Espérance mathématique (French), etc. However, the use of *E* is inconvenient because this is often used to denote an event and, in any case, it can hardly remain if one seeks to unify it with **P**. It is difficult to foresee whether this unification will command widespread support and lead to a search for terms with initial letter *P* in other languages (*see footnote above*), or other solutions. We say this to note that the proposed modification causes little difficulty in Italy, not only because of the existence and appropriateness of the term 'Previsione' but also because the international symbol *E* has not been adopted there.

In the present exposition we shall often, for the sake of brevity, use incorrect language, saying, for example: 'let the probability of E *be* $= \frac{1}{2}$', 'let the events A and B *be* (stochastically) independent', 'let the probability distribution of a random quantity X *be* normal', and so on. This is incorrect, or, more accurately, it is meaningless, unless we mean that it is a question of an abbreviated form to be completed by 'according to the opinion of the individual (for example You) with whom we are concerned and who, we suppose, desires to remain coherent'. The latter should be understood as the constant, though not always explicitly stated, intention and interpretation of the present author.

This is stated, and explicitly repeated, wherever it seems necessary, due to the introduction of new topics, or for the examination of delicate points—perhaps even too insistently, with the risk, and near certainty, of irritating the reader. Even so, notwithstanding the present remark (even imagining that it has been read), I am afraid that the very same reader when confronted with phrases like those we quoted, instead of understanding implicitly those things necessary in order to interpret them correctly, could have the illusion of being in an oasis – in the 'enchanted garden' of the objectivists (as noted at the end of Chapter 7, 7.5.7) – where these phrases could constitute 'statements' or 'hypotheses' in an objective sense.

In our case, in fact, the consequences of the pitfalls of the language are much more serious than they are in relationship to the Copernican system, where, apart from the strong psychological impediments due to man's egocentric geocentrism, it was simply a question of choosing between two objective models, differing only in the reference system. Much more serious is the reluctance to abandon the inveterate tendency of savages to objectivize and mythologize everything;[16] a tendency that, unfortunately, has been, and is, favoured by many more philosophers than have struggled to free us from it.[17] This has been acutely remarked, and precisely with reference to probability, by Harold Jeffreys:[18]

> 'Realism has the advantage that language has been created by realists, and mostly very naïve ones at that; we have enormous possibilities of describing the inferred properties of objects, but very meagre ones of describing the directly known ones of sensations.'

---

16 The main responsibility for the objectivizationistic fetters inflicted on thought by everyday language rests with the verb 'to be' or 'to exist', and this is why we drew attention to it in the exemplifying sentences by the use of italics. From it derives the swarm of pseudoproblems from 'to be or not to be', to 'cogito ergo sum', from the existence of the 'cosmic ether' to that of 'philosophical dogmas'.

17 This is what distinguishes the acute minds, who enlivened thought and stimulated its progress, from the narrow-minded spirits who mortified it and tried to mummify it: those who took every achievement as the starting point to presage further achievement, or those, on the contrary, who had the presumption to use it as a starting point on which to be able to base a definitive systematization.

For the two types, the qualification given by R. von Mises seems appropriate (see *Selected Papers*, Vol. II, p. 544): 'great thinkers' (like Socrates and Hume) and 'school philosophers' (like Plato and Kant).

18 Jeffreys, a geophysicist, who as such was led to occupy himself deeply with the foundations of probability, holds a position similar in many aspects to the subjectivistic one. The quotation is taken from H. Jeffreys, *Theory of Probability*, Oxford (1939), p. 394.

1.12 References

1.12.1. We intend to limit the present references to a bare minimum. The reader who wishes to study the topics on his own can easily discover elsewhere numerous books and references to books. Here the plan is simply to suggest the way which I consider most appropriate for the reader who would like to delve more deeply into certain topics, beyond the level reached here, without the inconvenience of passing from one book to another, with differences in notation, terminology and degree of difficulty.

1.12.2. The most suitable book for consultation according to this plan is, in my opinion, that of Feller:

Willy Feller, *An Introduction to Probability Theory and its Applications*, in two volumes: I (1950) (2nd and 3rd edn, more and more enriched and perfected, in 1956 and 1968); II (1966); John Wiley & Sons, Inc., New York.

The treatment, although being on a high level and as rigorous as is required by the topic, is not difficult to read and consult. This is due to the care taken in abolishing useless complications, in making, as far as possible, the various chapters independent of each other while facilitating the links with cross-references, and in maintaining a constant interplay between theoretical questions and expressive examples. Further discussion may be found in a review of it, by the present author, in *Statistica*, **26**, 2 (1966), 526–528.

The point of view is not subjectivistic, but the mainly mathematical character of the treatment makes differences of conceptual formulation relatively unobtrusive.

1.12.3. For the topics in which such differences are more important, that is those of inference and mathematical statistics (Chapter 11 and Chapter 12), there exists another work that is inspired by the concepts we follow here. Such topics are not expressly treated in Feller and thus, with particular reference to these aspects, we recommend the following work, and above all the second volume:

Dennis V. Lindley, *Introduction to Probability and Statistics from a Bayesian viewpoint*, in two volumes: I, *Probability*; II, *Inference*; Cambridge University Press (1965).

Complementing the present work with those of Feller and Lindley would undoubtedly mean to learn much more, and better, than from this work alone, except in one aspect; that is the coherent continuation of the work of conceptual and mathematical revision in conformity with the criteria and needs already summarily presented in this introductory chapter.

The above-mentioned volumes are also rich in interesting examples and exercises, varied in nature and difficulty.

2

Concerning Certainty and Uncertainty

2.1 Certainty and Uncertainty

2.1.1. In almost all circumstances, and at all times, we all find ourselves in a state of uncertainty.

Uncertainty in every sense.

Uncertainty about actual situations, past and present (this might stem from either a lack of knowledge and information, or from the incompleteness or unreliability of the information at our disposal; it might also stem from a failure of memory, either ours or someone else's, to provide a convincing recollection of these situations).

Uncertainty in foresight: this would not be eliminated or diminished even if we accepted, in its most absolute form, the principle of determinism; in any case, this is no longer in fashion. In fact, the above-mentioned insufficient knowledge of the initial situation and of the presumed laws would remain. Even if we assume that such insufficiency is eliminated, the practical impossibility of calculating without the aid of Laplace's demon would remain.

Uncertainty in the face of decisions: more than ever in this case, compounded by the fact that decisions have to be based on knowledge of the actual situation, which is itself uncertain, to be guided by the prevision of uncontrollable events, and to aim for certain desirable effects of the decisions themselves, these also being uncertain.

Even in the field of tautology (i.e. of what is true or false by mere definition, independently of any contingent circumstances), we always find ourselves in a state of uncertainty. In fact, even a single verification of a tautological truth (for instance, of what is the seventh, or billionth, decimal place of π, or of what are the necessary or sufficient conditions for a given assertion) can turn out to be, at a given moment, to a greater or lesser extent accessible or affected with error, or to be just a doubtful memory.

2.1.2. It would therefore seem natural that the customary modes of thinking, reasoning and deciding should hinge explicitly and systematically on the factor *uncertainty* as the conceptually pre-eminent and determinative element. The opposite happens, however: there is no lack of expressions referring to uncertainty (like 'I think', 'I suppose', 'perhaps', 'with difficulty', 'I believe', 'I consider it as probable', 'I think of it as likely', 'I would bet', 'I'm almost certain', etc.), but it seems that these expressions, by and large, are no more than verbal padding. The solid, serious, effective and essential part of arguments, on the other hand, would be the nucleus that can be brought within the

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

language of certainty – of what is certainly *true*, or certainly *false*. It is in this ambit that our faculty of reasoning is exercised, habitually, intuitively and often unconsciously.

In reasoning, as in every other activity, it is, of course, easy to fall into error. In order to reduce this risk, at least to some extent, it is useful to support intuition with suitable superstructures: in this case, the superstructure is *logic* (or, to be precise, the *logic of certainty*).

Whether it is a question of traditional verbalistic logic, or of mathematical logic, or of mathematics as a whole, the only difference in this respect is in the degree of extension, effectiveness and elegance. In fact, it is, in any case, a question of ascertaining the *coherence*, the *compatibility*, of stating, believing, or imagining as hypotheses some set of 'truths'. To put it in a different way: thinking of a subset of these 'truths' as *given* (knowing, for instance, that certain facts are true, certain quantities have given values, or values in between given limits, certain shapes, bodies or graphs of given phenomena enjoy given properties, and so on), we will be able to ascertain which conclusions, among those of interest, will turn out to be – on the basis of the data – either *certain* (certainly true), or *impossible* (certainly false), or else *possible*. The qualification 'possible' – which is an intermediate, generic and purely negative qualification – is applied to everything that does not fall into the two extreme limit cases: that is to say, it expresses one's ignorance in the sense that, on the basis of what we know, the given assertion could turn out to be either true or false.

2.1.3. This definition of 'possible' itself reveals an excessive and illusory confidence in 'certainty': in fact, it assumes that logic is always sufficient to separate clearly that which is determined (either true or false), on the basis of given knowledge, from that which is not. On the contrary (even apart from the possibility of deductions which are wrong, or whose correctness is in doubt), to the sphere of the *logically possible* (as defined above) one will always add, in practice, a fringe (not easily definable) of the *personally possible*; that is that which must be considered so, since it has not been established either that it is a consequence of one's knowledge or that it is in conflict with it.

We have already said, in fact, that logic can *reduce the risk of error*, but cannot eliminate it, and that tautological truths are not necessarily accessible. However, in order not to complicate things more than is required to guard against logical slips, we will always consider the case in which 'possible' can be interpreted as *logically possible*.[1]

2.2 Concerning Probability

2.2.1. The distinction between that which, at a certain moment, we are ignorant of, and that which, on the other hand, turns out to be certain or impossible, allows us to think about the range of *possibility*; that is, the range over which our uncertainty extends. However, this is not sufficient as an instrument and guide for orientation, decision or action: to this end –and this is what we are interested in – it will be necessary to base oneself on a further concept; the concept of *probability*.

---

1 Possibly by *eliminating* some knowledge. For instance, in the case of π it seems reasonable (for the problem under consideration) to imagine that one ignores the properties that permit the calculation of π, and to consider it as an 'experimental constant' whose decimal representation could only be known if somebody had determined it and published the result. I believe that for a mathematician, too, it would be reasonable to think that everything proceeds as if he were in such a state of ignorance.

In this chapter we do not wish to talk about probabilities, however; they will be introduced in Chapter 3. This deferment is undoubtedly awkward: obviously, the awkwardness consists in introducing preliminary notions without, at the same time, exhibiting their use. Didactically this is a bad mistake – one runs the risk of making boring and dull that which otherwise would appear clear and interesting. However, when it is important to emphasize an essential distinction, which otherwise would remain unnoticed and confused, a rigid separation is necessary – even if it seems to be artificial and pedantic. This is precisely the case here.

2.2.2. The study of the range of possibility, to which we shall here limit ourselves, involves learning how to know and recognize all that can be said concerning uncertainty, while remaining in the domain of the logic of certainty; that is, in the domain of what is *objective*. Probability will be a further notion not belonging to that domain and, therefore, a *subjective* notion. Unfortunately, these two adjectives anticipate a question about which there could be controversial opinions – their use here is not intended to prejudice the conclusion, however. For the time being, what matters is to make clear a distinction that is methodologically fundamental: afterwards, one can discuss the interpretation of the meaning of the two fields it delineates, the choice of nomenclature, and the points of view corresponding to them. It is precisely in order to be able to discuss them lucidly *afterwards* that it is necessary to avoid an immediate discussion of possibility and probability together; the confusion so formed would be difficult to resolve.

Both the distinction and the connection between the two fields are easily clarified: the logic of certainty furnishes us with the range of possibility (and the 'possible' has no gradations); probability is an additional notion that one applies within the range of possibility, thus giving rise to gradations ('more or less probable') that are meaningless in the logic of certainty.

2.2.3. Since it is certain that everyone knows enough about probability to be able to interpret these explanations in a less vague fashion, we can say that 'probability is something that can be distributed over the field of possibility'. Using a visual image, which at a later stage might be taken as an actual representation, we could say that the logic of certainty reveals to us a space in which the range of possibilities is seen in outline, whereas the logic of the probable will fill in this blank outline by considering a mass distributed upon it.

There is no harm in anticipating the developments that the treatment will undergo from the next chapter onwards, provided that, from the fact that they are not talked about here, one understands that they do not belong in the domain that we now consider it important to present as well-delimited and distinct.

2.3 The Range of Possibility

2.3.1. *Prologue*. Let us introduce right away the use of 'You,' following Good (Savage uses 'Thou'). The characterization of what is *possible* depends on the state of information. The state of information will be that (at a given moment) of a real individual, or it might even be useful to think of a fictitious individual (as an aid to fixing ideas). This individual, real or fictitious, in whose state of information – and, complementarily, of *uncertainty* – we are interested, we will denote by 'You.' We do so in order that You, the reader, can better identify

yourself with the rôle of this character. This character – or, better, You – will play a much more important rôle after this chapter, when probabilities will enter the scene. For the moment, You are in the audience, because You have to limit yourself to passively recording what You know for certain, or what You do not know.[2] All the same, it will be useful for You to at least get used to putting yourself in this character's place, since, even if it is not yet time to speak our lines, we are about to walk onto the stage – that is to enter into the range of possible alternatives.

With regard to any situation or problem that You have to consider, there will always exist an enormous number of conceivable alternatives. Your information and knowledge will, in general, permit You to exclude some of them as impossible: that is, they will permit You – and this has been said to be the function of science – a 'limitation of expectations'. All the others will remain *possible* for You; neither certainly true, nor certainly false. It will not happen that only one of them will be isolated as *certain*, except in special cases, or unless a rather crude analysis of the situation is given. Obviously, it is always sufficient to take all the possible alternatives and present them as a whole in order to obtain a single alternative which is 'certain'.

The choice of which of the more or less sophisticated, detailed, particularized forms we need, or consider appropriate, in order to distinguish or subdivide such alternatives, according to the problems and the degree of refinement we require in considering them, depends on us, on our judgment. Also, we have available several possible languages in which we can express ourselves in this connection. It is convenient to introduce them straight away, and altogether, in order to show, at the same time, on the one hand their essential equivalence, and, on the other, the differences between them which render their use more or less appropriate in different cases.

2.3.2. *Random events and entities*. Everything can be expressed in terms of *events* (which is the simplest notion); everything can be expressed in terms of *random entities* (which is the most generic and general notion); and so on. One or other of these notions is sufficient as a starting point to obtain all of them. However, it is instructive to concentrate attention on four notions which immediately allow us to frame within the general scheme the most significant types of problems, important from both the conceptual and practical points of view.

We will consider:

random events,
random quantities,
random functions,
random entities.

Let us make clear the meaning that we give to 'random': it is simply that of 'not known' (for You), and consequently 'uncertain' (for You), but *well-determined* in itself. Not even the circumstance of 'not known' is to be taken as obligatory; in the same way we could number constants among functions, though we will not call a constant a 'function' if there is no good reason. To say that it is *well-determined* means that it is *unequivocally individuated*. To explain this in a more concrete fashion; it must be specified in such a way that a possible bet (or insurance) based upon it can be decided without question.

---

2 You would have a more personal and autonomous rôle if we took into account the faculty, which You certainly possess, of considering as 'possible' that which You could show to be impossible, but which demands too much deductive effort. However, we have stated, in Section 2.1.3, that, for the sake of simplicity, we omit consideration of such hypotheses.

2.3.3. First, let us consider *random quantities*: this is an intermediate case from which we can pass more easily to the others, particularizing or generalizing as the case may be. We will denote a number, considered as a random quantity, by a capital letter; for example X or Y, and so on. It might be an integer, a real number, or even a complex number; but the latter case should be specified explicitly. The true value is unique, but if You call it random (in a nonredundant usage) this means that You do not know the true value. Therefore, You are in doubt between at least two values (possible for You), and, in general, more than two – a finite or infinite number (for instance, all the values of an interval, or all the real numbers). We will denote by $I(X)$ the set of possible values of X, and we will write, in abbreviated form, inf X and sup X for inf $I(X)$ and sup $I(X)$. It is particularly important to distinguish the cases of random quantities which are *bounded* (*from above and below*), that is inf X and sup X finite, and those which are only *bounded from above*, or only *bounded from below*, or *unbounded*, that is inf $X = -\infty$, or sup $X = +\infty$, or both.

To exemplify what we mean by *well determined* in the case of random quantities, let us put X = the year of death of Cesare Battisti.[3] The true value is $X = 1916$. While he was alive this value was not known to anyone and all years from that time on were possible values (for everybody). After the event, it is only random for those who are ignorant of it: for instance, for those who know only that it happened during Italy's participation in the World War I, the possible values are the four years 1915, 1916, 1917 and 1918.

Every function of a random quantity, $Y = f(X)$ or of two (or more), $Z = f(X, Y)$, and so on, is a random quantity (possibly 'degenerate', i.e. certain, if, for instance, $f(X)$ has the same value for all possible values of X).

2.3.4. An *event* (or *proposition*) admits only two values: TRUE and FALSE. In place of these two terms it is convenient to put the two values 1 and 0 (1 = TRUE, 0 = FALSE); in this way we simply reduce to a special case of the preceding, with an obvious, expressive meaning. Thus, when we wish to interpret the convention in this way, the event is identified with a gain of 1 if the event occurs and with a gain of 0 if the event does not occur. Moreover, with this convention the logical calculus of the events is simplified.

We continue to denote events with capital letters; in the main, $E, H, A, B, \ldots$. It is clear, for instance, that $1 - E$ is the *negation of E*, which is false if E is true, and vice versa (value 0 if $E = 1$, and conversely): it is also clear that AB is the logical product of A and B, that is true if both A and B are true, and so on (this is merely an example, the topic will be developed later, in Section 2.5).

An event corresponds to a question which admits only two answers; YES or NO (YES = 1, NO = 0). It is clear that with a certain number of questions of this type we can obtain an answer to a question that involves any number of alternative answers. Given a *partition* into s alternatives (one, and only one, of which is true), we can consider, for instance, the s events (exclusive and exhaustive) which correspond to them. But even less is sufficient: with n events we can imagine 2^n dispositions of YES–NO answers; we therefore have a partition into $s = 2^n$ alternatives if all these answers are possible, or into a smaller number, $s < 2^n$, if some of them are impossible (see Section 2.7 for further details).

Abandoning the restriction to a 'finite number', it is clear that by means of events we can study every case, even those involving an infinite number of possibilities.

---

3 Cesare Battisti was deputy for Trento at the Vienna Parliament; he volunteered for the Italian army, was then taken prisoner and hanged by the Austrians in 1916.
(Trento, where the author once lived, is an Italian city which was, in Battisti's time, a part of Austria.)

2.3.5. By talking about *random entities* in general, we have a means of expressing in a synthetic form the situation presented by any problem whatever. It is a question of referring oneself at all times to the same perspective, the one already implicitly introduced in the case of a random quantity, and which we now wish to make more precise and then to extend.

In the case of a random quantity, X, we can visualize the situation by considering as the '*space of alternatives*', $\mathscr{S}$, a line, the x-axis,[4] and on it the set, $\mathscr{Q}$, of the only values (points) *possible* (for You). In this way we consider *en masse*, implicitly, all the events concerning X (that it belongs to a half-line, $X \leqslant x$, or to an interval, $x' \leqslant X \leqslant x''$, or to any arbitrary set, $X \in I$).[5]

But now it is obvious that the same representation holds in all cases (in a more intuitive sense, of course, in three, or fewer, dimensions). If we consider two random quantities, X and Y, we can think of the Cartesian plane, with coordinates x and y, as the space $\mathscr{S}$ in which we have a set $\mathscr{Q}$ of *points* (pairs of values for X and Y) *possible* (for You) for a random *point* (X, Y). Every event (proposition, statement) concerning X and Y corresponds to a set I of $\mathscr{S}$: of course, only the intersection with $\mathscr{Q}$ is required, but it is simpler (and innocuous) to think of all sets I. The same could be said in the case of three random quantities X, Y, Z (in this case $\mathscr{S}$ is ordinary space), or for more than three.

Independently of the coordinate system, we could, in this geometric representation, formulate a problem straightaway. It might concern a *random point* on a plane (e.g. that point which would be hit in firing at a target), or in ordinary space (e.g. the position, at a given instant, of a satellite with which we have lost contact). We find an appropriate representation for the situation of a particle (position and velocity) by using six-dimensional space: the space of dimension $6n$ serves as 'phase space' for the case of n particles.

Independently of the geometrical meaning, or any meaning that suggests (in a natural way) a geometrical representation, we can always imagine, for any *random entity*, an abstract space $\mathscr{S}$ consisting of all possible alternatives (or, if convenient, a larger space of which these form a subset $\mathscr{S}$). We could consider, for example, *random vectors*, *random matrices* or *random functions*, and, thus far, the linear structure of the space continues to present itself as natural. But we could also consider *random sets*: for example, *random curves* (the path of a fly, or an aeroplane), random sets on surfaces (that part of the earth's surface in shadow at a given instant, or on which rain fell in the last 24 hours); or we could think of random entities inadequate to give any structure to the space.

We can, therefore, accept this representation as the general one, despite some reservations which will follow shortly (the latter are intended not as arguments against the representation, or for its rejection, but rather in favour of its acceptance 'with a pinch of salt').

---

4 We always denote by x(y, etc.) the axis on which X(Y, etc.) is represented.

5 We omit here critical questions relating to the possibility of giving, or not giving, a meaning to statements of an extremely delicate or sophisticated nature (or at least to the possibility of taking them into consideration). For example, the distinction between $<$ and $\leqslant$, the case of I 'nonmeasurable' in some sense or other, etc. It will be necessary to say something in Chapter 6; discussion of a critical character will be developed only in the Appendix, apart from brief anticipatory remarks here and there.

2.3.6. There is no need to deal with *random functions* separately, by virtue of the particular position they hold with respect to the preceding considerations (just as events and arbitrary entities have extreme positions, and random quantities an intermediate, but instrumentally fundamental, position). It is useful, however, to mention them explicitly for a moment. Firstly, in order to point out an example of applications which become more and more important from now on, and are largely new with respect to the range of problems traditionally recognized. Secondly, because we can allude, in a simple and intuitive way, to certain critical observations of the kind that will be reserved, in general, for the Appendix.

A *random function* is a function whose behaviour is unknown to You: we will denote it by $Y(t)$, assuming for convenience of intuition that the variable t is time.[6] If the function is known up to certain parameters, for instance $Y(t) = A\cos(Bt + C)$ with A, B, C random (i.e. unknown to You), the whole thing is trivial and reduces to the space of parameters. The case which, in general, we have in mind when we speak of a random function – or a *random process*, if we wish to place more emphasis on the phenomenon than on the mathematical translation – is that in which (to use the suggestive, if somewhat vague, phrase of Paul Lévy) the uncertainty exists at every instant (or, in his original expression, 'chance operates instant by instant').

This might mean, for example, that knowing the values of $Y(t)$ at any number of instants, $t = t_1, t_2,\ldots, t_n$, however large the (finite) n, the value at a different instant t will still, in general, be uncertain. Sometimes, either for simplicity or in order to be 'realistic', we imagine that it makes sense to measure Y at a finite (although unrestrictedly large) number of instants, without disposing of other sources of knowledge.[7] In such cases, the space $\mathscr{S}$ can be thought of as that in which every function is a 'point', but in which the possibility of distinguishing whether or not a function belongs to a set is only possible for those sets defined by a finite number of coordinates: the latter, being observable, are actually events. The simplest form of these events occurs when we ask whether or not the values at given instants fall inside fixed intervals $a_h \leqslant Y(t_h) \leqslant b_h$, $h = 1, 2,\ldots, n$. To give a visual interpretation, we ask whether or not the graph passes through a sequence of n 'doors', like a *slalom*.

2.4 Critical Observations Concerning the 'Space of Alternatives'

2.4.1. Having reference to the 'space of alternatives' undoubtedly provides a useful overall visualization of problems. Nevertheless, the systematic and, in a certain sense, indiscriminate use of it, which is fashionable in certain schools of thought, does have its dangers. One should learn to recognize these, and strive to avoid them.

In considering fields of problems of whatever complexity – in which, for instance, random sets, functions, sequences of functions and so on can occur together – the most

6 Our preference for $Y(t)$, rather than the more usual $X(t)$ as a notation for a generic random function, depends mainly on the fact that an X is often used as an 'ingredient' in the construction of $Y(t)$ At other times, x is used as a variable in place of t, and, anyway, in the graphical representation it is always convenient to think of the ordinate as y, and the abscissa as t or x.

7 Like, for instance, velocity $Y'(t)$ at an instant, measured with a speedometer; or the maximum or minimum of $Y(t)$ in an interval (t', t''), measured with instruments like a Max–Min thermometer.

general way of interpreting and applying the concepts exhibited in Section 2.3.5 is always the same; that is the following.

One goes back to the finest possible partition into 'atomic' events – not themselves subdivisible for the purposes of the problem under consideration – and these are considered as *points* constituting the set $\mathcal{Q}$ of 'possible outcomes'. This abstract space is the 'space of alternatives', or the 'space of outcomes': in certain cases, such as the examples of Section 2.3, it may be convenient to think of it as embedded in a larger and more 'manageable' space, and to regard this latter as the 'space of alternatives'.

In this scheme of representation, each problem (by which we mean problem concerning the alternatives $\mathcal{Q}$) reduces to considering 'the *true* alternative' (or 'the one which will turn out to be verified', or however one wants to express it), as a *random point* in $\mathcal{S}$ or, if we wish to be precise, in $\mathcal{Q}$. Let us call this point Q: it expresses everything there is to be said. Were we to lump together in $\mathcal{S}$ all possible problems, this space would be the space of all possible histories of the universe (explained as far as the most unimaginably minute details), and Q would be that point representing the true history of the universe (explained as far as the most unimaginably minute details).

Each event in this scheme is evidently interpretable as a set of points. E is the set of all points Q for which E is true; for example, it is the set of all individual 'histories of the universe' in which E turns out to be true. With the interpretation 1 = TRUE, 0 = FALSE, one could also say that E is a function of the point Q with values 1 on points Q of the set E, and 0 elsewhere (the indicator[8] function of the set E).

Similarly, each random quantity is interpretable as a real-valued function of the points Q: $X = X(Q)$ is the value which X assumes if the *true* point is Q. The preceding case, $E = E(Q)$, is simply the particular case which arises when the function can only take on the values 0 and 1.

The same is true for random entities of any other kind: for example, a random vector is a vector which is a function of the point Q.

2.4.2. That all this can be useful and convenient as a form of representation is beyond question. But things are useful if and only if we retain the freedom to make use of them when, and only when, they are useful, and only up to the point where they continue to be useful. A scheme that is too rigid, too definitely adopted and taken 'too seriously', ends up being employed without checking the extent to which it is useful and sensible, and risks becoming a *Procrustean bed*.

This is what happens to those who refer themselves too systematically to this scheme. Pushing the subdivision as far as the 'points' perhaps goes too far, but stopping it there creates a false and misleading dichotomy between the problems belonging, and not belonging, to the field under present consideration. The logical inconvenience which this already creates in the range of possibility will become far more dangerous and insidious when probabilities are introduced into such a structure.

An analogy between events and sets exists, but it is nothing more than an analogy. A set is effectively composed of elements (or points) and its subdivision into subsets

8 In a different terminology, the indicator function is also called the characteristic function: this term has many other meanings, and, in particular, in the calculus of probability it has a different and very important meaning for which it must be reserved (see Chapter 6).

necessarily stops when subdivision reaches its constituent points. With an event, however, it is possible, at all times, to pursue the subdivision (although in any application it is convenient to stop as soon as the subdivision is sufficient for the study in progress, otherwise things get unnecessarily complicated). The elements of the 'final subdivision' we have interpreted as 'points', but any idea which does not take into account the relative, arbitrary and provisional nature of such a delimiting of the subdivision, which thinks of it as 'indivisible', or as 'less subdivisible', or in any way different from all other events, is without foundation and misleading. For instance, it would be illusory to wish to distinguish between events corresponding to 'finite' or 'infinite' sets, or belonging to finite or infinite partitions, as if this had some intrinsic meaning. There is even less justification for retaining, as necessary, topological properties which happen to be meaningful in $\mathscr{S}$. The latter we referred to as 'space', instead of 'set', simply to use a more expressive language, and also because topological structures often exist and have interest in certain spaces by virtue of the nature of the spaces themselves, even when not required for any reason pertaining to the logical or probabilistic meaning.

2.4.3. Other objections, which we will develop a little more in the Appendix, would lead us to impugn even more radically the validity of the above representative scheme (and of many other things that we have hitherto admitted and which, for the moment, we continue to admit). As an example, we note the fact that all sets (or the 'points' of them) must be accepted as having the meaning of events.

In general terms, it will always be a question of examining if, and in which sense, a statement really constitutes an 'event', permitting, in a more or less realistic and acceptable form, and in a unique way, the 'verification' of whether it is 'true' or 'false'.

What should be said concerning statements that are 'verifiable' only by means of an infinite number of observations, or by waiting an infinite length of time, or by attaining an infinite precision? A critical attitude in this respect could lead one not to consider as 'events' the fact that X has exactly the value x, or belongs to a set of measure zero (e.g. is rational), but only the fact that $X \in I$ for a set I 'up to sets of measure zero' (and this, although it eliminates some difficulties, introduces others), or 'up to an error $<\delta$, that can be chosen as small as desired, but nonzero', and so on. Even more radical are the difficulties of 'complementarity', which appeared first in quantum physics but can be detected on a smaller scale in more everyday examples: A and B are events (observable), but it is not possible to observe both of them, and, therefore, it is not possible to call the product AB an event (observable).

All this, in addition to the specific reasons already given in the main text (and to which we return in the next paragraph), reduces the value of the reduction to 'points'. Indeed, it is symptomatic that, precisely in connection with arguments of this kind, von Neumann developed a 'geometry without points' (in 'Continuous geometries', *Proc. Nat. Acad.*, **22** (1936), 92–100 and exemplified *Proc. Nat. Acad.*, **22** (1936), 101–108) where, as he says: 'The point which we wish to stress is that the investigations described above show an unbroken trend *away from the notion of the point*'. The studies to which he alludes are those of K. Menger and G. Bergmann (on linear spaces), of F. Klein, G. Birkhoff and O. Ore (on lattices), and discussions with J.W. Alexander and H. Veblen.

Even more strictly in accordance with the considerations in the text, appear to be the studies of St Ulam (in the 'von Neumann lecture', Princeton (1963), still unpublished), since he also refers himself to structures *open* to the adjunction of new entities as new

circumstances arise. A 'continuous geometry' of von Neumann, on the other hand, is a closed structure, although very rich, containing linear systems of any dimension *c*, with *c* any real number between 0 and 1 (the empty and complete systems, respectively). Ulam says: The indications are … that *there are no atoms of simplicity* and, which is most strange, one would almost be tempted to say that in the physical world the set-theory *axiom of Regularity* – that is to say, that *every set contains a minimal element with respect to the relation of "belonging to a set" – does not hold*!'.[9]

2.5 Logical and Arithmetic Operations

2.5.1. Having, through the convention 1 = TRUE, 0 = FALSE, given to events an interpretation that makes them particular random quantities, it becomes both possible and useful to take advantage of this unification in order to effect also an appropriate unification of the operations related to them. Usually, and inevitably, prior to such a convention,[10] one considers two distinct series of operations: the (Boolean) *logical operations*

$$\wedge \textit{ logical product}; \quad \vee \textit{logical sum}; \quad \sim \textit{negation}$$

applicable only to *events*; and the *arithmetic operations*

$$\cdot \textit{ product}; \quad + \textit{sum} \,(\text{and their inverses}: / \text{ and } -)$$

applicable only to *numbers*.

We have already touched upon the utility of certain applications of the arithmetic operations to events, automatically possible by the above convention (see Section 2.3.4, and also allusions in Chapter 1). We are now able not only to develop this extension systematically, but also to obtain a complete unification by extending, in the opposite direction, the logical operations into the field of numbers.

In fact, in the field of (real) numbers, we make the definitions:

$$x \wedge y = \min(x, y), \quad x \vee y = \max(x, y), \ \sim x = 1 - x \,(= \tilde{x}).^{11}$$

It is immediate that the definitions agree with those known in the field of events (that is, of the idempotent numbers 0 and 1), whereas, obviously, the usual properties (which it would be beneficial to interpret and understand through examples in each of the two cases), always hold both for numbers and events:

---

9 The italics are present in the original for the last three words only.

10 Which, as I later discovered, had already been adopted by von Neumann in 1932 in his treatment of quantum mechanics; Appendix, Section 9.

11 As usual, we agree to place the tilde for 'complementary to 1' above, instead of in front, when dealing with a single letter. The same convention – using a bar rather than a tilde – was adopted by L. Dubins and L.J. Savage, *How to Gamble if You Must*, McGraw-Hill (1965), p. 64, and found to be of frequent utility.

$$\left.\begin{aligned} \sim(x \wedge y) &= \tilde{x} \vee \tilde{y} \\ \sim(x \vee y) &= \tilde{x} \wedge \tilde{y} \end{aligned}\right\} \text{ (duality of } \wedge \text{ and } \vee \text{ with respect to complements),}$$

$$\left.\begin{aligned} x \wedge (y \vee z) &= (x \wedge y) \vee (x \wedge z) \\ x \vee (y \wedge z) &= (x \vee y) \wedge (x \vee z) \end{aligned}\right\} \text{ (distributivity between } \wedge \text{ and} \vee \text{),}$$

$$\left.\begin{aligned} x \wedge x &= x \\ x \vee x &= x \end{aligned}\right\} \text{ (idempotence for } \wedge \text{ and } \vee \text{)}$$

(in addition to the obvious commutative and associative properties of $\wedge$ and $\vee$).

2.5.2. *Operations on events*. By virtue of what has already been said, it is not a question of making new definitions, but only of applying the general definitions to the case of the values 0 and 1; it remains only to establish agreement with the usual meaning.

By the *logical product* of two (or more) events A, B, we mean the event which is true if and only if all the factors are, and therefore false if at least one is false. If the factors can only be 0 and 1, both the arithmetic product and the operation *min* ($\wedge$) obviously enjoy the property that the result is 1 if and only if all the factors are 1. Therefore, *in the field of events*, the two operations of arithmetic product and logical product coincide; thus, we could always refer simply to the *product* of two events, without danger of ambiguity, and write $E = AB$. The symbol $\wedge$ might be used for greater clarity only in complicated cases; for instance,

$$E = \left(X + Y \geqslant 54\right) \wedge \left(Z \geqslant Y + 12\right),$$

where the events are conditions (on random quantities X, Y, Z etc.), written as parentheses, and the fact that they are events and not numbers could be overlooked.

By the *negation* of an event A, we mean the event that is true if A is false and vice versa; obviously, we have 'not A' = $\sim A = \tilde{A} = 1 - A$, because $\sim 1 = 1 - 1 = 0$, $\sim 0 = 1 - 0 = 1$.

By the *logical sum* of two (or more) events A, B, we mean the event that is true if at least one of the summands is true, and therefore false if and only if they are all false. To this corresponds the operation *max* ($\wedge$), which gives 1 if at least one summand is 1, and 0 if all summands are 0. It is also obvious and well known that, with respect to negation, the operation is dual to that of the product:

$$A \vee B = \sim\left(\tilde{A} \wedge \tilde{B}\right).$$

This follows also from the properties stated generally for $\tilde{x} \wedge \tilde{y}$.

This allows us to obtain an arithmetic expression for the logical sum: taking complements and expanding, we obtain

$$\mathrm{A} \vee B = 1 - (1 - A)(1 - B) = A + B - AB, \tag{2.1}$$

and, similarly,

$$\begin{aligned} \mathrm{A} \vee B \vee C &= 1 - (1 - A)(1 - B)(1 - C) \\ &= A + B + C - AB - AC - BC + ABC. \end{aligned}$$

In general, for n summands,

$$E_1 \vee E_2 \vee \ldots \vee E_n = \sum_i E_i - \sum_{ij} E_i E_j + \sum_{ijh} E_i E_j E_h - \ldots \pm E_1 E_2 \ldots E_n, \tag{2.2}$$

where the sums have to be taken over all the n events E_i over all the $\binom{n}{2}$ products two at a time, over all the $\binom{n}{3}$ products three at a time, and so on, with alternate signs, up to the last term which is the product of all n events with + if n is odd, – if n is even.

The *arithmetic sum* of two (or more) events A, B, is not, in general, an event, but a random number expressing the *number of successes*. In particular, $A + B$ has either the value 0 (if they are both false), or 1 (if one is true and the other false), or 2 (if they are both true). In general, as in this case, the relation between logical sum and arithmetic sum is the following: *both have the value* 0 *if every summand happens to be false* (no successes), whereas, otherwise, if true summands (successes) exist and number 1, 2, 3,..., in general *m*, *the* (*arithmetic*) *sum is that number*, whereas the *logical sum* always takes the value *one*; that is, does not take into account multiplicity,

$$(\text{logical sum}) = 1 \wedge (\text{arithmetic sum}) \tag{2.3}$$

or, explictly,

$$E_1 \vee E_2 \vee \ldots \vee E_n = 1 \wedge \left(E_1 + E_2 + \ldots + E_n\right). \tag{2.3'}$$

The fact of having two distinct notions is not, therefore, inconvenient but, on the contrary, is an advantage because both have their raison d'être. We are still faced with the problem of eliminating the ambiguity of the terminology – since we do not wish to be obliged to say 'logical sum' or 'arithmetic sum' every time. For this purpose it is sufficient to adopt the natural convention of using *sum* for the arithmetic sum, and *event-sum* for the logical sum (because only this is an event).

2.5.3. We observe that the operations introduced induce, over the field of real numbers, the structure of a *lattice*, with the operation ~ which enjoys many properties of the complement (in the algebraic sense), but is not exactly such, except in the field of events (the numbers 0 and 1). There, in fact, we have $x \vee \tilde{x} = 1$ (because either x or $\tilde{x}$ is 1, and the other 0), in addition to $x + \tilde{x} = 1$, which is also valid for any x.

In addition, we observe that the expressions in arithmetic form for $\sim x$, $x \wedge y$, $x \vee y$ coincide (in the field of events) with those of Stone, where the sum has to be taken 'mod 2', however, in order to obtain a Boolean ring.

The conventions adopted here do not give rise to algebraic properties of this kind but seem to be the most suitable for expressing, simply and naturally, many things which are otherwise difficult to express.

We will give examples at the end of this chapter (Section 2.11) in order not to interrupt the flow of the argument, and we will often use similar simplifications. It will be seen that it is not only a question of expressions concerning events or random quantities: for identical reasons, the same conventions meet requirements which also occur in other fields.

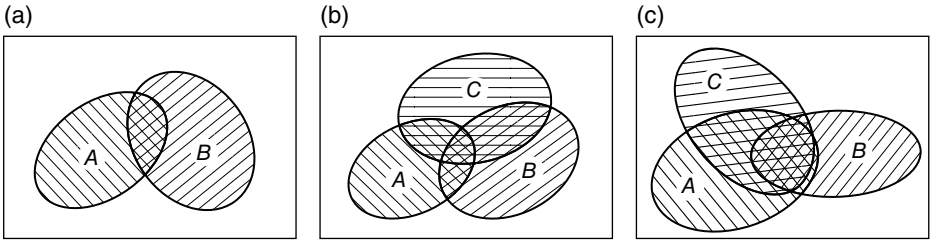


Figure 2.1 Venn diagrams: the representations of events and their logical relationships in the set-theoretic interpretation: (a), (b) the cases of two and, respectively, three events with all (4 and, respectively, 8) constituents possible; (c) an example in which only six of the eight combinations give (possible) constituents.

2.5.4. We have mentioned, in Section 2.3, the set-theoretic interpretation. It is clear that, by interpreting the events as sets, the operations $\sim$, $\wedge$, $\vee$, which we have introduced, correspond in that context to the set-theoretic operations $\sim$, $\cap$, $\cup$ (*complementation*, *intersection*, *union*). For random quantities, understood as functions of the 'point' Q, $Z = X \vee Y$ is the function that, at each point Q, assumes the larger of the two values $X(Q)$ and $Y(Q)$:

$$Z(Q) = X(Q) \vee Y(Q) \quad \text{(and similarly for } \wedge\text{)}. \tag{2.4}$$

A geometrical representation (which is formally identical) is especially useful, particularly for didactic purposes, even if a genuine set-theoretic interpretation is lacking: it is that of the so-called 'Venn diagrams'. The events which one wishes to represent are drawn as areas of a rectangle, which itself represents the certain event. The areas are delimited with lines or, better, distinguished with different types of shading. In this way, one can illustrate visually the relationships that are supposed to exist among the different events: the existence, or not, of a certain intersection – distinguished by the overlapping of different shadings – the inclusion of one event in another; and so on. Of course, it is only in rather simple examples that clear figures, whose areas are not too contorted, are possible.

Shown in Figure 2.1a and 2.1b are the cases of two and three events, respectively, where all the four (or eight) intersections are nonempty, that is are possible events; whereas in Figure 2.1c two of the pair-wise intersections are not present.

2.6 Assertion, Implication; Incompatibility

2.6.1. We began this chapter by saying that, for You, every event, or proposition, can be either certain, or impossible, or possible. We then talked about possibility. The time has now come to translate these premises into a precise argument. We must make a distinction that, in the terminology proposed by B.O. Koopman,[12] could be called a distinction between *contemplated propositions* and *asserted propositions*. As considered so far, a proposition E is always a contemplated proposition (for which You, or anyone else,

12 *The Bases of Probability*, in Kyburg and Smokler, pp. 161–172.

could know whether it is true or false). Thus it remains, even if changed into $E = 1$, or $\sim E = 0$, or $(E = 1) = 1$, and so on, or, put into words, 'E is true', 'not-E is false', 'it is true that E is true'. Nothing is altered, because these are simply more or less extended ways of saying nothing more and nothing less than E.

To make an *assertion*, we have to step outside of the vicious circle by saying something extra-logical; such as 'I assert that E is true', 'For You, it is certain that E is impossible', 'For me, E is possible': that is something expressing not a logical relationship between propositions, but a relationship between the proposition and the *speaker*.

To denote this succinctly, the symbol $\vdash$ has been introduced. If E is a proposition, an event, then, by using $\vdash$ as a prefix, $\vdash E$ becomes the *assertion* that 'E is certain' (for someone). Naturally, $\vdash \sim E$ is the assertion that 'E is impossible', whereas by $\sim \vdash E$ we mean to denote the assertion that 'E is possible' (i.e. the nonassertion of both E and of not-E).

2.6.2. We shall not make much use of this symbol, because we think that, in general, the distinction will be clear from the context (for instance, by saying 'certainly'). It is useful, however, to draw attention to the importance of the distinction, and to illustrate the use of the symbol by giving some examples in order to fix all this in the reader's mind. In any case, these observations were necessary at this juncture in order to make it clear that certain expressions, which we will now introduce, *have to be taken as assertions*.

By saying that an event A *implies* the event B, or that A is *contained* in B, we mean to *assert* that A cannot occur unless B also occurs, or that $A\tilde{B}$ is impossible: in symbols $\vdash \sim A\tilde{B}$. Instead of $\sim A\tilde{B}$ one may also write $\tilde{A} \vee B$ or $A\tilde{B} = 0$, or $A \leqslant B$, or $B - A \geqslant 0$ (because the inequality is false only for $1 \leqslant 0$, i.e. for $A = 1$ and $B = 0$). It is always a question of ways of expressing $\sim A\tilde{B}$, independently of the fact that it is certain, or impossible, or possible, and these give assertions, simply by making the assertions. In order to write that 'A implies B' with the meaning, as we have said, of assertion, it will be necessary to write, for example, $\vdash A \leqslant B$. However, we will introduce some ad hoc symbols, to be understood as already having the value of assertions:

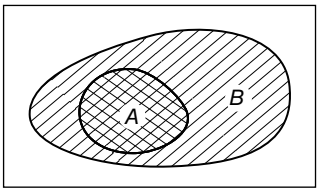
$A \subseteq B. =. \vdash A \leqslant B,$	A implies B;
$A \equiv B. =. \vdash A = B,$	A is identical to B (or $A \subseteq B \wedge B \subseteq A$), or, A and B are either both certainly true or both certainly false: (*certain*) *equality of A and B;*
$A \subset B. =. A \subseteq B \wedge \sim A \equiv B,$	A strictly implies B.[13]

13 The *equality*, $A = B$, is the event that takes place if A and B are both true or both false, and this can happen for any A and B (except in the case of complementary events, $B = \tilde{A}$). However, in order not to make the language unnecessarily heavy, we will continue to say, as usual, 'equal', rather than 'certainly equal', and to write $=$, rather than $\equiv$, except in ambiguous cases.

As regards *strict implication*, observe that it asserts that $A \leqslant B$ with certainty, but that $A = B$ is not certain. In other words, we exclude $A > B$, i.e. A true and B false, but we do not exclude the converse, $A < B$. Nothing is said concerning the possibility or impossibility of A and B being either both true, or both false. Observe that $A \subset B$ means $(A \subseteq B) \wedge \sim(A \equiv B)$ or $(\vdash A \leqslant B) \wedge (\sim \vdash A = B)$, which is very different from $\vdash \left[(A \leqslant B). \sim (A = B)\right] = \vdash A < B = \vdash \tilde{A}B$, which denotes the assertion that A is false and B is true.

The meaning of all these relationships is immediately, intuitively obvious under the set-theoretic interpretation.

Figure 2.2 Venn diagram: the case of implication (inclusion).



2.6.3. The relationship of implication, which is clearly reflexive and transitive, induces, over any set of events, a partial ordering and in particular a lattice in which the operations $\wedge$ and $\vee$ (in the sense of 'maximal' element contained in those given, and 'minimal' element containing the given ones) coincide with those of logical product and logical sum, already introduced. This is evident above-all under the set-theoretic interpretation $A \subseteq B$ means that *A is a subset* of *B*, possibly coincident with *B* (this being excluded if we write $A \subset B$, affirmed if we write $A \equiv B$); hence the terms 'contains', 'is contained in', have the opposite meaning to 'imply', 'is implied by', instead of being synonymous as they might appear to be if one thought, for both terms, of the interpretation in terms of events.[14] In other words, in the Venn diagram for two events, which 'in general' (more precisely, for *A* and *B* logically independent) has the appearance of Figure 2.1, the part of *A* not contained in *B* must be missing (empty); in other words, *A* must coincide with the doubly shaded area *AB* (as in Figure 2.2).

If both regions with single shading are missing we have the case $A \equiv B$, and if the other two regions (double shading and no shading) are missing we have $A \equiv \tilde{B}$. Two other important cases correspond to the absence of the doubly shaded area (case of incompatibility: $AB \equiv 0$), or the absence of the nonshaded area (case of exhaustivity: $\tilde{A}\tilde{B} \equiv 0$).[15]

2.6.4. *Incompatibility*. By saying that two events *A* and *B* are *incompatible*, we mean to *assert* that it is impossible for them both to occur; i.e. that *AB* is impossible: in symbols $\vdash \sim AB$. Instead of $\sim AB$ we can write $AB = 0$, or $\tilde{A} \vee \tilde{B}$, or $A + B = A \vee B$, or $A + B \leqslant 1$, or $A \leqslant \tilde{B}$, or $B \leqslant \tilde{A}$, always expressing the event $\sim AB$, independently of the fact that it is certain or impossible or possible. Each of these forms expresses the incompatibility; if it is asserted, we can write, e.g., $\vdash A + B \leqslant 1$, or $\vdash A \leqslant \tilde{B}$, which can be expressed, by reduction to the implication, as $A \subseteq \tilde{B}$. By saying that *n* events $E_1, E_2, \ldots, E_n$ are incompatible, we mean to assert that they are pairwise incompatible ($\vdash E_i E_j = 0$, $i \neq j$); that is that at most one of them can occur. As a straightforward extension of $\vdash A + B \leqslant 1$, this can be expressed as $\vdash Y \leqslant 1$, where $Y = E_1 + E_2 + \ldots + E_n$ is the number of 'successes'; that is of the events E_i that are true. The same definition also holds for an infinite number of events: in this case, instead of a non-negative integer, *Y* could also be an infinite cardinal (e.g. that of denumerability, or of the continuum, or any other aleph). We note also that the condition $E_1 + E_2 + \ldots + E_n \equiv E_1 \vee E_2 \vee \ldots \vee E_n$, that is the coincidence of the logical and arithmetic sums,[16] is always characteristic of the case of incompatibility.

---

14 To avoid possible consequent mnemonic uncertainties about the meaning of $\subseteq$ (and hence the opposite meaning for $\supseteq$), it is sufficient to think of it as corresponding to $\leqslant$ ($\supseteq$ then corresponds to $\geqslant$), whose meaning is clear if we consider operations on the numbers 0 and 1 (events, indicator functions of sets).

15 The other cases are trivial: *A* or *B* or both would be determined, either certain or impossible.

16 For any non-negative (random) numbers the same conclusion is obviously valid: such equality holds if and only if at most one of them can be nonzero.

In other words, incompatible events are mutually exclusive; in the set-theoretic interpretation it is a question of *disjoint* sets, having an *empty intersection* (pairwise, and hence, *a fortiori*, for three or more).

2.6.5. *Exhaustivity*. By saying that two events A and B are *exhaustive*, we mean to *assert* that it is impossible for neither of them to occur; i.e. that $\tilde{A}\tilde{B}$ is impossible: in symbols $\vdash\sim \tilde{A}\tilde{B}$. Instead of $\sim \tilde{A}\tilde{B}$ one can (as above) write $\tilde{A}\tilde{B}=0$. or $\boldsymbol{A \vee B}$, or $\tilde{A}+\tilde{B}=\tilde{A}\vee\tilde{B}$ (i.e. $2-(A+B)=1-AB$, $A+B=1+AB$), or $A+B\geqslant 1$, or $\tilde{A}\leqslant B$ or $\tilde{B}\leqslant A$; another form for the exhaustivity is therefore, for instance, $\vdash A+B\geqslant 1$. This lends itself easily to the extension of the definition to the case of n events, or even to an infinite number. By saying that these are exhaustive (or, better, form an exhaustive family – but the phrase is cumbersome), we mean to assert that at least one of them must take place; that is, in the preceding notation, $\vdash Y\geqslant 1$. This shows the relationship between the two conditions. In the set-theoretic interpretation, it is a question of a family of sets which *covers* the whole set Q of possible points (of course, there may be some overlapping); i.e. those sets of points for which the complement of the union is empty.[17]

2.7 Partitions; Constituents; Logical Dependence and Independence

2.7.1. *Partitions*. A *partition* is a family of *incompatible and exhaustive* events – that is for which it is *certain* that one and only one event occurs. The coexistence of the conditions $\vdash Y\leqslant 1$ and $\vdash Y\geqslant 1$ means, in fact, $\vdash Y=1$. A partition can be finite or infinite: partitions (and, for the simplest conclusions, in particular finite partitions) have a fundamental importance in the calculus of probability (which, as already indicated, will consist in distributing a unit 'mass' of probability among the different events of each partition).

It is, therefore, of importance to see now if, and how, one can reduce the general case, in which one considers any finite number of events $E_1, E_2,\ldots, E_n$, to that of a partition. We observe first of all that if, in particular, the E_i are already incompatible, but not exhaustive, it will be sufficient to add on the extra event

$$E_0=1-\left(E_1+E_2+\ldots+E_n\right)\quad\left(\text{i.e., in another form, } E_0=\tilde{E}_1\tilde{E}_2\ldots\tilde{E}_n\right).$$

In the general case, we must consider the 2^n products $E_1'E_2'\ldots E'$ where each E_i' is either E_i, or its complement $\tilde{E}_i$; formally, we can obtain them as the individual terms of the expansion $(E_1+\tilde{E}_1)(E_2+\tilde{E}_2)\ldots(E_n+\tilde{E}_n)$, which is identically 1, since each factor is 1. Some of the 2^n terms may turn out to be impossible and do not have to be considered: those which remain, and are therefore possible, are called the *constituents* $C_1, C_2, \ldots, C_s$ of the partition determined by $E_1, E_2, \ldots, E_n$, where $s\leqslant 2^n$.

---

17 Suppose that, instead of considering Q – the space of possible points for You, now – one considers a larger space S which contains, in addition, certain points that are already known to be impossible (for instance, in the light of more recent information). In all the preceding cases, the statement that a set is *empty* must be replaced by *empty of possible points* – i.e. empty of points belonging to Q. In diagrams, one could think of the region $S\sim Q$ as drawn in black, and consider it as 'nonexistent'.

By observing that the given expansion has value 1, we have already established that we are dealing with a partition; on the other hand, the fact is evident per se (even more so under the set-theoretic interpretation). A partition is given by a family of disjoint sets that covers the space Q; or, in other words, into which Q is subdivided – in the same way, for instance, in which Italy is divided into municipalities. If, instead, we perform any other division whatsoever, the partition given by the constituents is that into the 'pieces' resulting from such a subdivision. For instance, Italy east and west of the Monte Mario meridian, north and south of a given parallel, areas of altitude above and below 500 metres, areas more or less than 50 kilometres from the sea, belonging to a province the name of whose capital or main city begins with a vowel or consonant, and so on.[18]

Sometimes it will also be useful to introduce the (clumsy) notion of a '*multi-event*' for cases in which (provided we do not restrict ourselves to meaning 'event' in a purely technical sense) a partition might correctly be called an 'event with many alternatives'. Such is a game – a football game, for instance – with the three alternatives 'victory', 'draw' and 'defeat' (and possibly a fourth, 'not valid' because of postponement etc.). The same holds in the case of drawings from an urn containing balls of three or more different colours, for example 'white', 'red', 'black'; or throwing a die, or two dice, with possible points in the range 1–6, or 2–12, respectively. A multi-event with m alternatives – more briefly an '*m-event*' – can always be thought of as a random quantity with m possible values (e.g. 1, 2,..., m). In the case of a single die, the 'points' are precisely 1, 2,..., 6, whereas for the two dice it is irrelevant whether we use 2, 3,..., 12, or 1, 2,..., 11. The colours, or results of the game, could similarly be coded numerically. In speaking of an m-event we want, essentially, to emphasize the *qualitative* aspects of the alternatives. It is then appropriate to use the mathematical interpretation of them as unit vectors (1, 0,..., 0), (0, 1, 0,...,0),..., (0, 0, 0,..., 1) in an m-dimensional space. In this way, writing E_h (h = 1, 2,...,m) for the events[19] which consist in the occurrence of the hth alternative, an m-event can be identified with the random vector (E_1, E_2,..., E_m). The (arithmetic) sum of multi-events gives, therefore, the number of occurrences of the single results: for instance, (W, R, B) = the number of drawings of White, Red and Black balls. We observe the analogy with the case of events, which could be handled in this same way, by substituting (1, 0) for 0 and (0, 1) for 1 (if the advantage of the symmetry seemed to compensate for the unnecessary introduction of the doubleton).

2.7.2. *Logical dependence and independence of events.* We define n events (necessarily *possible*) to be *logically independent* when they give rise to 2^n possible constituents. This means that each of these events remains uncertain (possible) even after the outcomes of all the others, whatever they may be, are known: this explains the choice of terminology. In fact, let us suppose that one of the products is impossible, and therefore only a constituent in a formal sense – without loss of generality, take it to be $E_1E_2 \ldots E_n$. E_1 is possible, E_1E_2 may or may not be, and the same holds for $E_1E_2E_3$, $E_1E_2E_3E_4$, and so on. If one of these products is impossible, obviously all the subsequent ones are; the last one – the

18 Caution: do not think of separate parts of a unique nonconnected 'piece' as 'pieces' – the topology of the representation must be ignored. The 'piece' of Italy north-east of Monte Mario with altitude below 500 metres and more than 50 kilometres from the sea in a province beginning with a vowel is certainly composed of separated parts (for instance in the provinces of Ancona and Udine).

19 Necessarily incompatible and exhaustive.

product of all the n events – is impossible by hypothesis, and therefore either it or one of the preceding ones must be the first to be impossible: suppose this is $E_1E_2E_3E_4$. This means that it is possible for events E_1, E_2 and E_3 to occur and that, knowing this, we are in a position to exclude the possibility that E_4 can be true. The events are therefore in this case – that is if the number of constituents is $s < 2^n$ – logically dependent.

Of course, if n events are logically independent the subsets of the n are, *a fortiori*, independent: the converse does not hold. Even if all their proper subsets exhibit logical independence n events can still be logically dependent. As a simple example, let all the constituents in which the number of successes is even be possible, and no others; this imposes no restrictions on the result of any $n - 1$ events whatsoever, but for each event the result is determined once we know the results of the others.

2.7.3. If one wishes to consider more specifically the dependence of a *particular event* E on certain others, $E_1, E_2,\ldots, E_n$, it becomes necessary to consider several cases. It is, in fact, possible that E remains uncertain after we know the results of the Ei, whatever these may be: we then call it *logically independent*. On the other hand, it is possible that it will always be determined (either true or false), in which case we call it *logically dependent*. However, an intermediate case could also arise: the uncertainty or the determination of E might depend on the actual results of the E_i; this we will call *logical semidependence*. We could be more precise and refer to logical semidependence *from below*, or *from above*, or *two-sided*, according to whether there exist outcomes for the E_i which make E *certain*, or *impossible*, or whether there exist outcomes of both types.

In order to characterize the various types of event, with respect to the fixed E_i, it suffices to consider the constituents determined by the E_i. We have $C_1 + C_2 + \ldots + C_s = 1$, and each event E can, therefore, be decomposed into $E = EC_1 + EC_2 + \ldots + EC_s$. For any one of the summands, say EC_h, there are three possibilities: either $EC_h = C_h$ (if C_h is contained in E), or $EC_h = 0$ (if C_h is contained in $\tilde{E}$), or else $0 \subset EC_h \subset C_h$ (if both EC_h and $\tilde{E}C_h$ are possible). The possible results for the E_i correspond to the occurrence of one of the constituents C_h: according to whether C_h is of the first, second or third type, E turns out to be *certain*, *impossible* or remains *uncertain*, respectively.

The conclusions are obvious.

E is *logically dependent* if constituents of the third type do not exist; that is if E is a sum of constituents (of the first type). We could also say that E is logically dependent on the E_i if and only if it is expressible as a function of them by means of logical operations: in this case we have dependence by definition. The value (true or false) of such an expression is, in fact, determined by the values of the variables appearing in it; conversely, every such expression reduces to a *canonical form* as a sum of constituents and, therefore, the condition is also necessary. In this case, constituents of both the first and second types exist; otherwise, E would have been either certain or impossible to begin with, contrary to hypothesis.

E is *logically independent* if all the constituents are of the third type, and *logically semidependent* if some, but not all, are of the third type: in the latter case, we have semidependence *from below* if the others are all of the first type, *from above* if they are all of the second type, *two-sided* if there are some of each type.

If we consider the two events

E' = the sum of all the constituents of the first type, and
E'' = the sum of all the constituents of the first and third types,

it clearly turns out that in each case $E' \subseteq E \subseteq E''$, and E' and E'' are, respectively, the maximal event certainly contained in E, and the minimal event that certainly contains it – that is the events giving the best possible bounds.

We can then say that: E is logically dependent if $E' = E''$ (and hence $= E$); logically independent if $E' \equiv 0$ and $E'' = 1$; semidependent from below, from above, or two-sided, if

$$0 \subset E' \subset E'' \equiv 1, \quad 0 \equiv E' \subset E'' \subset 1, \quad 0 \subset E' \subset E'' \subset 1,$$

respectively.

2.7.4. These notions of logical dependence and independence are meaningful more generally; they apply not only to the case of events, as considered so far, but also to partitions, or random quantities, or any random entities whatsoever. We will present the development for the case of random quantities, which is the most intuitive; it will then suffice to remark that the concept is always the same.

Random quantities (suppose, to fix ideas, that there are three: X, Y, Z) are said to be *logically independent* if there are no circumstances in which the knowledge of some of them can modify the uncertainty concerning the others. This means that if X, Y, Z have, respectively, r possible values x_i, s possible values y_j, t possible values z_h, then all the rst triples (x_i, y_j, z_h) are possible for (X, Y, Z); that is the set Q of possible points (x, y, z) is the *Cartesian product of the sets*, Q_x, Q_y, Q_z of possible values for X, Y, Z. In this form the definition is general: it is valid not only for n (instead of 3), but also if the random quantities have an infinite number of possible values (for instance, those of an interval), or in the case of random entities of other kinds, or, generically, for partitions,[20] In other words, the condition means that nothing, no known interdependence, allows any further restriction of the set Q of possible points over and above that resulting from the fact that the individual random quantities, or entities, must assume values in $Q_1, Q_2, \ldots, Q_n$.

2.7.5. The *logical dependence* of one (random) quantity on others (to fix ideas consider the dependence of Z on X and Y) has exactly the meaning that it has in analysis: Z is a *function* (i.e. a *one-valued* function) of X and Y, $Z = f(X, Y)$, the function $z = f(x, y)$ being defined for all the possible points (x, y) of (X, Y). *Logical independence* means that the set of possible values of Z conditional on the knowledge of the values of X and Y (any pair of possible values (x, y) for (X, Y)) is always the set of Q of all the (unconditionally) possible values of Z. Intermediate cases, which are not worth listing in further detail, always give *logical semidependence*.

2.7.6. A critical observation is appropriate at this point, both as a refinement of the present argument and to exemplify various cases in which it is useful to examine whether the logic needs to be taken with a pinch of salt (see Appendix).

We will confine ourselves to a single example. Suppose that X and Y have as possible values all the numbers between 0 and 1, with the condition that $X + Y$ is irrational : then Q is the unit square with an infinite number of 'scratches' removed – parallel to the diagonal and corresponding precisely to the lines $x + y =$ rational. For the partition into

20 A partition can be reduced to a random quantity by considering as such, for example, the index i of Ei:$X = i$ if Ei is true, provided the partition has at most the cardinality of the continuum, or is denumerable if we require integer values.

points, logical independence does not hold; it would hold, however, for every partition into vertical or horizontal stripes, however small.

Is it advantageous to say that we do not have logical independence when its failure is attributable to subtleties of this kind? Clearly, there is no categorical answer. It seems obvious that, depending on the problem and on one's intentions, one decides whether or not to take such subtleties into account (of course, one must be careful to be precise when such is required).

2.7.7. Finally, we make an observation which, strictly speaking, is unnecessary – being implicit in the very definition of 'possibility' – but which it is convenient to make and underline. All the notions we have encountered, or introduced – from incompatibility to logical dependence or independence – are relative to a given 'state of information'. They are valid for You (or for me, or him) according to the knowledge, or the ignorance, determining the uncertainty; that is the extent of the range of the *possible*, of the set Q (yours, mine, his,…).

It is a question of relative and personal notions, but nonetheless objective, in the sense that they depend on what one knows, or does not know, and not on one's opinion concerning what one does not know, and what is, consequently, uncertain.

In order to avoid ambiguity, we must never forget that we are always speaking about uncertainty in the simple sense of ignorance. In particular, of course, we are dealing with matters traditionally attributed to 'chance' – a trace of this remains in the word 'random',[21] and in other expressions which we will be using. In general, however, we are concerned with any future matters whatsoever, and also of things in the past concerning which there is no information, or for which no information is available to You, or which You cannot remember exactly: we might even be concerned with tautologies. The various cases differ in one important aspect: that is the existence and degree of possibility and facility of obtaining, in one way or another, further information, should one wish to do so. This fact will, of course, be relevant in determining behaviour in decision problems, where it could be convenient to condition on the acquisition of new information. But apart from this, basically, it is convenient to regard any distinctions of this kind as unimportant. The only essential element, which determines and characterizes our object of study, is the existence of imperfect information – of whatever kind – and the situations of uncertainty in which, consequently, You might find yourself.

There is a prejudice that uncertainty and probability can only refer to future matters, since these are not 'determined' – in some metaphysical sense attributed to the facts themselves instead of to the ignorance of the person judging them. In this connection, it is useful to recall the following observation of E. Borel: 'One can bet on Heads or Tails while the coin, already tossed, is in the air, and its movement is completely determined; and one can also bet after the coin has fallen, with the sole proviso that one has not seen on which side it has come to rest'.

2.7.8. *Remark*. It might be useful to point out (or, for those who already know it, to recall the fact) that in the theory of probability one often uses the term 'independence' (without further qualification) to denote a different condition, that of *stochastic independence*, which refers to *probability* and will be introduced in Chapter 4.

---

21 *Translators' note*. The Italian word here is '*aleatorio*' (see French, aléatoire) from the Latin *alea* meaning *die*: 'alea jacta est !' – the die is cast !' – as Caesar said when crossing the Rubicon.

Be careful not to confuse it with *logical* independence – which we have just discussed – or with *linear* independence, which we will discuss in the next section. Both of these notions have an objective meaning; that is independent of the evaluation of the probabilities.

2.8 Representations in Linear form

2.8.1. *Basic notions*. When referring to the set Q of possible 'points' in the case of two random quantities X and Y, we tacitly interpreted the pair (x, y) as Cartesian coordinates in the plane (which it was natural to take as the space of alternatives S). Similarly, for three or more points, we extend to ordinary space, or to spaces of any dimension (always in Cartesian coordinates).

This was simply a question of habit and, therefore, of convenience. One could have thought of any coordinate system; of a curved surface instead of a plane, or, in order to say more and in a better way, it is enough to think in terms of a space in a merely abstract sense, for which such distinctions of a geometric nature do not even make sense. With reference to the simplest case, it is sufficient that different pairs (x, y) are made to correspond to distinct 'points'.

For further reasons, which we now wish to take into account – because, as we shall see in Chapter 3, they are essential for the theory of probability – it becomes important instead to think of S as a *linear* (*affine*) *space*. We shall call it the *linear ambit* and denote it by A because at times it will be convenient to consider as the space S not the whole of A but a less extensive manifold which contains Q. For example: if A is ordinary space, and X, Y, Z are related by the equation $X^2 + Y^2 + Z^2 = R^2$, it might be convenient to think of S as the spherical surface on which one finds the possible points Q; these may consist of all the points of the surface, or a part of it, or just a few points, depending on other restrictions and circumstances and knowledge.

A representation that is *linear* with respect to certain random quantities (e.g. those considered initially) is such with respect to others that are linear combinations of them (but not with respect to the rest). If we require that linearity holds for the rest too, we have to extend the linear ambit A to new dimensions, as we shall see later.

The random quantities linearly represented in an ambit A themselves constitute a linear system, which we denote by S, and which is dual to A. One might ask whether it is useful to think of the two dual spaces, A and S, as superposed. In principle, the answer is no: in fact, only the affine notions have any meaning, and the metric, introduced surreptitiously by means of such a superposition, would be dependent on the arbitrary choice of the coordinate system that has to be superposed onto its dual. In general, for this reason, it is not even practically convenient. A unique exception is perhaps that of the case we considered first, in which we start from *events*, and it is 'natural' to represent them with unit, orthogonal vectors. In any case, whether or not this possibility is useful in a particular case, it is important never to forget that it is only the affine properties which make sense.

These properties also underlie the notions and methods fundamental to the theory of probability. On the other hand, the things in question are very elementary, and are currently applied without first introducing this formulation and terminology – which might well be considered excessively theoretical and, for the purpose in hand,

disproportionately so. Nevertheless, if one is prepared to make the small effort necessary to picture the question in terms of the present scheme, many aspects of what follows will appear obvious and well-connected among themselves, instead of, as they might otherwise appear, unrelated and confused. So much so, that the preeminent – one might even say exclusive – rôle of linearity in the theory of probability has always remained very much in the background. This is, in part perhaps, because of the prominence given to the Boolean operations, and because of the nonimmediacy of the arithmetic operations on events when the latter are not identified with their 'indicators'. The present treatment is intended to provide the framework within which these observations will find their justification and clarification.

2.8.2. Let us begin by considering events $E_1, E_2,\ldots, E_n$, and, often, in order to be able to think in terms of ordinary space, we will, without essential loss of generality, take n to be three.

The linear ambit A is the affine vector space in n dimensions, with coordinate system $x_1, x_2,\ldots, x_n$, in which we will consider the values of the random quantities $X_1, X_2, \ldots, X_n$. In this case, the latter are the events $E_1, E_2,\ldots, E_n$, taking only the values 0 and 1: the set of 'possible' points consists at most, therefore, of the 2^n points (8, if $n = 3$) with coordinates either 0 or 1, and may be a subset of these. One sees immediately – as was inevitable – that the 'possible' points correspond to the s ($s \leqslant 2^n$) constituents.

Given the special rôle of these points, it is convenient to think of the prism, of which they are the vertices, as a cube (or hypercube) and, therefore, to think of the Cartesian coordinate system x_i as orthogonal and of unit length – with the reservation that this metric not be taken too 'seriously'.

The linear system L, of linear combinations of $E_1, E_2,\ldots E_n$, consists of random quantities $X = u_1E_1 + u_2E_2 + \ldots + u_nE_n$,[22] interpretable as the *gain* of someone who receives an amount u_1 if E_1 is true, plus an amount u_2 if E_2 is true, and so on (of course, the 'gains' may be positive or negative). The X possess at most as many (distinct) possible values as there are constituents – namely s – and the latter occurs if the corresponding 'possible points' are found on distinct hyperplanes $\Sigma_i u_i x_i = \text{constant}$.

An important example is that where $Y =$ *the number of successes*. In order to obtain this, it is sufficient to take all the $u_i = 1$ – a gain of 1 for each event – obtaining, as we have already shown directly, $Y = E_1 + E_2 + \ldots + E_n$. In this case, it is clearly not true that the possible points occur on distinct hyperplanes; if all the 2^n vertices of the hypercube are possible, they are, in fact, distributed over the $n + 1$ hyperplanes $Y = 0, 1, 2,\ldots, n$ according to the binomial coefficients $(1, n, \frac{1}{2}n(n-1),\ldots, n, 1), \binom{n}{h}$ being the number of possible ways of obtaining h successes in n events.

For the case $n = 3$, we shall denote the Cartesian coordinates of the ambit A in the usual manner, by x, y, z, and those of the dual system L by $u, v\ w$. If $X = uE_1 + vE_2 + wE_3$, then $ux + vy + wz$ is the value which X would assume if E_1 takes the value x, E_2 the value y and E_3 the value z. Given the meaning of the E_i such values can only be either 0 or 1, and the value of the random quantity X (e.g. gain) can only be one of those corresponding to the eight vertices of the cube (or to a part of it, if not all the vertices are possible).

---

22 In order to simplify this example we omit the constant u_0 (see Section 2.8.3).

Here are the coordinates of such vertices Q together with the corresponding values of X:

$$\begin{array}{lccccccccc} Q & = & (0,0,0), & (0,0,1), & (0,1,0), & (1,0,0), & (0,1,1), & (1,0,1), & (1,1,0), & (1,1,1), \\ X & = & 0 & w & v & u & v+w & u+w & u+v & u+v+w. \end{array}$$

In particular, for $u = v = w = 1$, we see (as was obvious) that the number of successes is 0 in one case, 1 in three cases, 2 in three cases and 3 in one case. In addition (apart from the combinatorial meaning, $(1 + 1)^3 = 1 + 3 + 3 + 1 = 8$), this shows that, when projected onto a diagonal, the vertices of the cube fall as follows: one at each end, and three each at 📄 and 🤌 of the way along the diagonal.

2.8.3. The sum $\sum_i u_i x_i$ (in particular $ux + vy + wz$) is a linear function both of X (i.e. of its components u_i), and also of $\mathscr{Q}$ (i.e. of its coordinates x_i). We will denote it both by $X(Q)$ – thinking of it as 'the value of a given X as Q varies' – and also by $Q(X)$ – thinking of it as 'the value assigned to different X by the resultant Q' The same operation, however, will still turn out to be useful independently of the fact that Q is a possible point (i.e. $Q \in \mathscr{Q}$). That is, by replacing Q by any A in $\mathscr{A}$, writing $X(A)$ or $A(X)$:

$$A(X) = X(A) = \sum_i u_i x_i ,$$

where the u_i are the coordinates of the X considered as points of $\mathscr{L}$, or, better, the components of X considered as vectors of $\mathscr{L}$, and similarly the x_i are coordinates (or components) of the A considered as points (or vectors) of $\mathscr{A}$.[23] The expressions $A(X)$ or $X(A)$ then appear as *products* of vectors, A and X, belonging to the two dual spaces $\mathscr{A}$ and $\mathscr{L}$.[24]

What we have said so far in this section is independent of the assumption that, rather than taking any random quantities whatsoever, we start with events, $X_i = E_i$ (as we did in Section 2.8.2, in order to fix ideas). Since it is convenient to consider not only the homogeneous linear combinations, $X = \sum_i u_i X_i$, as we have up until now, but also complete combinations with an additional constant, say u_0, we will always assume as added to the X_i a fictitious random quantity X_0, taking the single value $X_0 \equiv 1$ *with certainty*. The summand $u_0 X_0$ has precisely the value u_0, with no alteration to the formula; we have only to take into account that there is an additional, fictitious, variable, x_0, and that, for all possible points (and, usually, also for every A to be considered), we will have $x_0 = 1$.

2.8.4. *Linear dependence and independence.* We have considered $\sum_i u_i X_i$ $(i = 0,1,2,\ldots n)$, linear combinations (either homogeneous or complete) of n random quantities X_i (i = 1, 2,..., n); X is said to be linearly dependent on the X_i. It may be, however, that the X_i are already linearly dependent themselves; that is that one of their linear combinations is identically zero (or constant: due to the inclusion of X_0 the two are essentially identical),

---

23 Given that the point O (the origin) has meaning in both L and A, there is no risk of ambiguity in identifying points and vectors.

24 If one thinks of the two spaces as superposed – we have already said that, in general, this is not advisable – we would have the scalar product. In any case, one could write AX and XA, instead of $A(X)$ and $X(A)$, thinking in terms of the product rather than writing it as a 'function'. The main application, however, will be when $A = P$ (probability, prevision), and the omission of the parentheses in this case – although used by some authors – seems to give less emphasis to the structure of the formulae, and therefore to the meaning.

in which case at least one of the X_i is a linear combination of the others and can be eliminated (because it already appears as a combination of the others). Geometrically, this means that the set Q of possible points belongs to a linear subspace A′ of A, and hence it is sufficient to confine attention to A′: the extension from A′ to A is illusory – one adds only points which are certainly impossible.

We observe that linear dependence is a special case of logical dependence – that is that linear dependence is a more restrictive condition. Conversely, it goes without saying that logical independence is more restrictive than linear independence.

We now return, briefly, to the case of events, for even here the distinction between linear dependence and logical dependence is of fundamental importance for the theory of probability. The negation of E depends linearly on E: in fact, $\tilde{E} = 1 - E$. On the other hand, the *logical product* $E = AB$, and the *logical sum* $E = A \vee B$, do not depend linearly on A and B (except when, under the assumption that A and B are incompatible, the logical sum has the form $A \vee B = A + B$). However, the logical sum does depend linearly on the two events and their product: $A \vee B = A + B - AB$. In general, the logical sum of three or more events depends linearly on the events themselves and on their products two at a time, three at a time,..., and finally the product of all of them (see Section 2.5.2). Apart from these cases of a general nature, however, it is possible that an event can be a linear combination of others 'by chance' (so to speak): an example can be found in Chapter 3, in connection with a probability problem, where an event E is expressed linearly as a function of others by the following formula

$$E = \tfrac{1}{7}\left(3 - 2E_1 + E_2 - E_3 + 3E_4 + 5E_5 - 5E_6\right).$$

How can one tell whether or not such a linear dependence exists? It is sufficient to express all events as sums of constituents and then to see whether the matrix (consisting entirely of zeroes and ones) is zero or not.

2.8.5. The above considerations refer to the system $\mathscr{S}$, but linear dependence is still meaningful and important in the ambit $\mathscr{A}$. The interest there lies in considering the barycentre P of two points Q_1 and Q_2 with 'masses' q_1 and q_2, where $q_1 + q_2 = 1$. By a well-known property in mechanics – which is, on the other hand, an immediate consequence of linearity – each linear function X assumes at P the value $X(P) = q_1X(Q_1) + q_2X(Q_2)$, and the same holds for the barycentre of three, or (leaving ordinary space) any number of points whatsoever. The property even holds if some of the masses are negative, but the cases in which we are normally interested are those with non-negative masses (usually, in fact, we will be dealing with probability).

The barycentre can, therefore, be any point[25] belonging to the *convex hull* of the points Q_h under consideration. Consideration of the convex hull determined by the 'possible points', $Q \in \mathscr{Q}$ or, in other words, the convex hull of $\mathscr{Q}$ will play a fundamental rôle in the calculus of probability. Dually (and this property too, well-known and intuitive, will turn out to be meaningful in future applications), the convex hull is also the intersection of all

25 If the points Q_h are infinite in number, then in order for this to be true we must also allow 'limit cases' of barycentres (which, in other respects, correspond to actual requirements of the calculus of probability, at least according to the version we will follow, in which we do not assume 'countable additivity'). Anyway, apart from questions of interpretation, this simply means that by convex hull we mean the set of barycentres completed by their possible adherent points.

the half-spaces containing $\mathcal{Q}$ In other words, if a point P belongs to the convex hull $K(I)$ of a set I, then it is on the same side as I with respect to any hyperplane not cutting the set – that is which leaves it all on the same side. On the other hand, if a point does not belong to the convex hull, there exists a hyperplane separating it from I – that is which does not cut the latter and leaves it all on the opposite side with respect to the point. Translating all this into an analytic form: every non-negative linear function on I is also such on $K(I)$; conversely, the property does not hold for any point not belonging to $K(I)$.

2.8.6. Returning to the case of the cube (Section 2.8.2), we already have a meaningful example, although a little too simple, of the way in which the convex hull varies as we consider all eight vertices or a subset of them (see Chapter 3, where the probabilistic meaning will also appear).

With this example in mind, it is now possible to make an observation which, although trivial in this context, is useful for explaining in an intuitive way our immediate intentions (Section 2.8.7) in cases where it could seem less obvious and perhaps strange.

In the space $\mathscr{A}$ we could represent the eight constituents by the vertices of the cube: we suppose that all eight actually exist, there is no need to consider other cases here. In the dual space $\mathscr{L}$, however, we could only represent the random quantities depending linearly on E_1,E_2,E_3. The eight constituents, considered as random quantities, could not be represented, and so neither could the random quantities derived from them linearly – unless these happened to be linearly dependent on the three fundamental events E_i. Does the method create a discrimination between events which have a representation as vectors in $\mathscr{L}$ and those which do not? If so, can we put the situation right?

The answer to the first question is no: the method creates no discrimination. The fact is that it enables us to consider more or fewer dimensions according to what we need. The representation in terms of the cube is sufficient for the separation of the eight constituents (as points of $\mathscr{A}$), and for the consideration of random quantities linearly dependent on the three E_h. If we wished, we could even reduce to a single dimension by considering only the random quantity $X = 4E_1 + 2E_2 + E_3$: this is sufficient to characterize the eight constituents, since X can assume the values 0, 1, 2, 3, 4, 5, 6, 7. These values, incidentally, are obtained by reading the triple of coordinates as a binary number – for example (1, 0, 1) = 101 (binary) = 4 + 0 +1 = 5. If we were interested only in such an X (up to linear transformations, $aX + b$), and in distinguishing the constituents, this would be sufficient. Similarly, if in addition to X we are interested, for instance, in the number of successes $Y = E_1 + E_2 + E_3$, and nothing else, we could pass to two dimensions. Suppose, however, that, for reasons which depend on the linearization, we are interested in studying, in $\mathscr{L}$, either one of the constituents, or a linear combination of constituents not reducible to a linear combination of the E_h. In this case, it will be necessary to introduce a third dimension and then, if required, others…, up to seven. In general, if there are s constituents we require $s - 1$ dimensions (s if we include a fictitious one for the constant $X_0 \equiv 1$) *in order that everything geometrically representable in $\mathscr{A}$ is also linearly interpretable in $\mathscr{L}$.*

In fact, if in our case (that of the cube) we consider an eight-dimensional space, whose coordinates x_h give the value of the constituents C_h, the possible points, Q_h, are the points with abscissa 1 on one of the eight axes (because one, and only one, of the eight constituents must occur). They are linearly independent in the seven-dimensional space $x_1 + x_2 + \ldots + x_8 = 1$: one of the x_h is superfluous, but it makes no difference whether we leave it, or eliminate it and add a fictitious coordinate $x_0 \equiv 1$. In terms of $\mathscr{L}$,

we can therefore obtain all the X either as linear combinations $\sum u_h C_h$, for h from 1 to 8, or for h from 0 to 7 (excluding C_8 but adding the fictitious $C_0 \equiv X_0 \equiv 1$).

Conclusion: everything can be represented linearly provided one takes a sufficient number of dimensions. It is possible, and this provides a simplification, to reduce this number by projecting onto a subspace (although in this way we give up the possibility of distinguishing between those things which have the same projection). Thus, for instance, different possible cases may be confounded into a single one, or even if we take care to avoid it, barycentres arising from different distributions of mass may be confounded. In the case of the cube, for example, each internal point can be obtained as the barycentre of $\infty^{7-3} = \infty^4$ different distributions of mass on the eight vertices.

2.8.7. In the general case, considering any random quantities whatsoever, the same circumstance arises and has even greater interest. Suppose we consider the ambit $\mathscr{A}$ relative to n random quantities X_i ($i = 1, 2,\ldots, n$) and, for simplicity, let us assume that all the real values are possible and compatible for the X_i: that is that all the points of $\mathscr{A}$ are possible ($\mathscr{A} = \mathscr{Q}$). It follows that every random quantity $Z = f(X_1, X_2,\ldots, X_n)$ is *geometrically* individuated in $\mathscr{A}$ (to each point of $\mathscr{A}$ there corresponds, in a known way, a value of Z), but is not *vectorially* represented in $\mathscr{L}$ unless it is a *linear* function of the X_i. If such a vectorial representation for Z is needed, however, it is sufficient to add on a new dimension for it – that is to introduce an extra axis, z, or, if one prefers, x_{n+1}, on which Z can be represented.

To give an intuitive illustration: in the plane (x, y) every function $z = f(x, y)$ already has a geometrical representation (visually through contour lines), but in order for z to appear *linearly* in the representation it is necessary to introduce a new axis, z, and to transfer each contour line to the corresponding height, obtaining the surface $z = f(x, y)$.

As a practical example, in fact one which continuously finds application, an even simpler case will suffice. We have a single random quantity, X: by taking the x-axis as the ambit $\mathscr{A}$, we represent, by means of its points, all the possibilities (values x) which determine, together with x, every function of x, $f(x)$. However, if we are interested in the linear representation of a given $f(x)$ we must introduce a new axis, y, and on it represent $y = f(x)$. The linear ambit A will be the plane (x, y), but for the space $\mathscr{S}$ we could more meaningfully consider the curve $y = f(x)$, whereas $\mathscr{Q}$ could be a set of points on such a curve (if not all values are possible for X). It will be, so to speak, the set $\mathscr{Q}$, previously thought of on the x-axis, projected onto the curve $y = f(x)$. We note, incidentally, that this illustrates the observation made in Section 2.8.1 regarding the nonidentification of A and S. The criterion which has been followed can be explained in the following way: we delimit S by taking into account the 'essential' circumstances, considering as such the fact of studying X together with a given $Y = f(X)$, whatever the random quantity X may be; we do not take into account the 'secondary' circumstances, considering as such the particular facts or knowledge which, in certain cases or at certain moments, lead us to exclude the possibility of X attaining certain values.

The most important practical case (which we have already mentioned) is the simplest one: that of X and $Y = f(X) = X^2$. The curve is the parabola $y = x^2$, and the linear system L consists of all the polynomials of second degree in X; $aX^2 + bX + c$. Suppose that we are interested in barycentres of possible points Q_h with given masses q_h. If the points are taken on the parabola we obtain a point $\overline{x}, \overline{y}$, which is meaningful for both coordinates, whereas if we leave the points on the x-axis the barycentre would give the same $\overline{x}$, but no information about $\overline{y}$.

Obviously, if we were interested in considering $Z = X^3$ also (i.e. extending L to polynomials of the third degree) it would be necessary to take the space (x, y, z) as the ambit A, the curve $y = x^2$, $z = x^3$ as the space S, and to project onto it the set Q already given; and so on.

2.9 Means; Associative Means

2.9.1. Within this representation, we will take the opportunity to present, in an abstract form, a notion which has great practical and conceptual importance in all fields, and which, in what follows, will above all prove useful in connection with probabilistic and statistical interpretations. The notion in question is that of a mean. This is usually defined in terms of mere formal properties of particular cases, but (as Oscar Chisini pointed out) it has a well-defined and important meaning as a useful 'summary' or 'synthetic characteristic' of something more complicated.

A prime example (already considered in the preceding pages) is that of the barycentre, or, arithmetically, that of the arithmetic mean (in general weighted) of the coordinates of the point masses. It is well known how, in mechanics, for many aspects and consequences, everything proceeds *as if* the whole mass were concentrated at the barycentre. In the language of statistics (which we will encounter mainly in Chapters 11 and 12) one would say that knowledge of the barycentre (and of the mass) constitutes, for certain purposes, a *sufficient statistic* (i.e. an exhaustive summary). For other purposes, in mechanics, it is necessary to know in addition the moments of inertia, and the exhaustive summary is then the collection of these items of information *of first and second orders*. It is convenient to point out in advance that knowledge of the second-order characteristics will also play an important rôle in statistics and in the theory of probability. Above all, it gives a powerful tool for studying problems in a way that is often sufficiently exhaustive, although summary.

2.9.2. Let us now consider the definition of mean according to Chisini, which is based precisely on this concept of an exhaustive summary. In this way we impart to the notion the *relative functional* meaning conveyed by '*tailor-made*' (better the German *Zweckmässig*, whose equivalent is missing in other languages: zweck = purpose, mässig = adequate). According to Chisini,[26] '*x is said to be the mean of n numbers* $x_1, x_2, \ldots, x_n$, *with respect to a problem in which a function of them* $f(x_1, x_2, \ldots, x_n)$ *is of interest, if the function assumes the same value when all the* x_h *are replaced by the mean value x:* $f(x_1, x_2, \ldots, x_n) = f(x, x, \ldots, x)$' Here we are considering the simplest case, without *weighting*, but the concept is still the same in the latter case, and in that – as we shall see in Chapter 6 – of *distributions*, even continuous ones.

2.9.3. The most important type of mean is the *associative* one. The defining property of associative means is that they are unchanged if some of the quantities are replaced by their mean (in the same way as, in order to find the barycentre, one can concentrate some of the masses at their barycentre). Independently, and almost simultaneously, Nagumo and Kolmogorov proved that the associative means are all, and only, the

26 O. Chisini, 'Sul concetto di media', in *Periodico di Matematiche* (1929); the topic is taken up again in an article by B. de Finetti in *Giorn. Ist. Ital. Attuari* (1931). The proof of the theorem of Nagumo and Kolmogorov can also be found there.

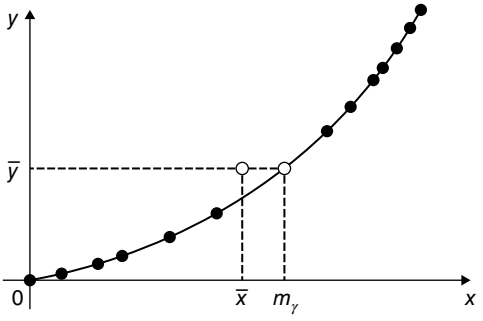


Figure 2.3 Comparison between associative (γ-) means based on comparisons of the convexity of the functions $\gamma(x)$ used to construct them.

(increasing) transforms of the arithmetic mean. They are obtained by taking an increasing function $\gamma(x)$, and, given the values x_h with respective weights p_h $\left(\sum p_h = 1\right)$, instead of taking the barycentre, $\overline{x} = \sum_h p_h x_h$, one takes the barycentre of the corresponding $y_h = \gamma(x_h)$, $\overline{y} = \sum_h p_h y_h$ and then reverts to the 'scale' x by means of the inverse function $m_\gamma = \gamma^{-1}(\overline{y})$ thus obtaining the γ-mean.

The procedure can be clearly 'seen' in the representation of the preceding paragraph. If we consider the example given there, we have $y = \gamma(x) = x^2$, and, of course, we must limit ourselves to the positive semi-axis in order that γ be increasing.[27] It is a question of thinking of the masses p_h as placed on the parabola; the barycentre is the point whose coordinates are $\overline{x}$ and $\overline{y}$, whereas $x = m_\gamma$ (obtained as shown in Figure 2.3) is the point to which corresponds (on the parabola) the same ordinate of the barycentre, the square root of the mean of the squares.

Considering the other function $z = x^3$ (either by itself in the plane (x, z), or together with $y = x^2$ in the space (x, y, z), as noted in Section 2.8.7), the barycentre would be $\overline{x}, \overline{z}$, respectively, $\overline{x}, \overline{y}, \overline{z}$, where $\overline{z}$ = the mean of the cubes = $\sum_h p_h x_h^3$, and $\sqrt[3]{\overline{z}}$ = the cube root of the mean of the cubes = the cubic mean of the values x_h with weights p_h, and so on. In Chapter 6 we will say something about the most important associative means: these correspond to $\gamma(x)$ = powers (with any positive or negative real exponent whatsoever; if zero we have the limit case of the logarithm), and exponential. At this point, however, it is convenient to consider some general properties related to the notion of convexity of which we have spoken. This will also clarify a few questions which we will meet in Chapter 3.

2.9.4. The barycentre is always in the convex polyhedron (or, in general, the convex hull) determined by the point masses: in our example we can think of it both in the plane and in ordinary space. For the main conclusion of interest to us, the case of the plane is sufficient. If the masses are on a curve whose concavity is always in the same direction, or on a portion of the curve for which this is true, the barycentre is always in the area bounded by the concavity; hence: *the γ-mean is greater than the arithmetic mean if γ (increasing) is concave upwards*. The quadratic mean is, therefore, greater than the arithmetic mean and so is the cubic mean: the question arises, can these two

27 Or the negative one. In fact, as is easily seen, $\gamma_1(x)$ and $\gamma_2(x)$ are equivalent with respect to the mean if (and only if) $\gamma_1 = a\gamma_2 + b (a \neq 0)$. If we change the sign of a (i.e. change increasing into decreasing) nothing is altered. It is clear from the diagram, in fact, that a change in y, either of scale or sign, or a vertical translation of the curve, makes no difference.

be compared? Of course; it is sufficient to project the curve $y = x^2$, $z = x^3$ (explicitly, $z = y^{3/2}$) onto the plane (y, z): the concavity is upwards and so the cubic mean is greater.

Even without the graphical comparison, it is sufficient to take into account that 'greater relative concavity' (in the sense that a diagram would display) corresponds, locally, to a greater value of $\gamma''(x)/\gamma'(x)$ (in the interval of interest if the function is not everywhere invertible). In the above example, we have $y''/y' = 2/2x = 1/x$, $z''/z' = 6x/3x^2 = 2/x$ and so, for $x > 0$, z''/z' is always greater. More generally, since for the powers $\gamma(x) = x^c$ one has $\gamma''/\gamma' = c(c-1)x^{c-2}/cx^{c-1} = (c-1)/x$, the mean increases with the exponent; this also holds for $\log x$ (the limit case as $c \to 0$: $\log x \cong (x^c - 1)/c$): in fact, $\gamma''/\gamma' = -x^{-2}/x^{-1} = -1/x = (0-1)/x$. This particular choice ($c = 0$, $\gamma = \log$) gives the geometric mean, which, in the case of two, or more generally n, values with equal weights (the 'simple', unweighted case) assumes the more familiar forms: $\sqrt{(x_1 x_2)}$, $\sqrt[n]{(x_1 x_2 \ldots x_n)}$, respectively. For $c = 1$, we have the harmonic mean, the reciprocal of the mean of the reciprocals.

From the fact that $-1 < 0 < 1 < 2 < 3$ it follows that in the above-mentioned cases, for example, we have:

harmonic < geometric < arithmetic < quadratic < cubic.

2.9.5. *Remarks*. Although it may seem strange to do so, we conclude by saying that the following observation is important: the barycentre of points which are on a curve (other than a straight line) is not a point on the curve – unless perhaps 'by chance'. In the same way, the barycentre of points on a surface (not a plane) is not, generally speaking, a point of the surface; and so on, in any dimension. The observation may seem strange because it is so obvious: its obviousness, however, results from the demonstration in terms of the above representation. How many people would recognize the fact before having their attention drawn to it? In facing real problems one often reasons as if what one considers strange, and even absurd, is precisely this fact!

2.10 Examples and Clarifications

2.10.1. Examples are always useful in order to give a sense of concreteness to concepts introduced in a general and abstract form. In this case, they will serve in addition to underline the meaning and importance of certain refinements, either already mentioned in passing or to be added soon, and also to introduce, before we yet talk about probability, a few of the kinds of situation which we will repeatedly come across in various problems.

Above all, by selecting widely differing examples we intend to remove any possible residual doubts that might lead to restrictive interpretations of the field of uncertainty to which we refer ourselves. The subject matter to which the uncertainty refers is irrelevant: political or economic events, meteorological phenomena, historical or scientific conjectures, judicial investigation, personal or everyday affairs, competition in sport, or any other field in which uncertainty and imperfect knowledge are present. This includes of course –and they in no way differ from the others – the traditional games of chance. This latter is, in fact, the least interesting case, because it leads to a standardized scheme in which all the conceptual and substantial aspects of the problem are made to disappear.

2.10.2. *Examples of events*. Will a given candidate, on a given occasion, succeed in getting elected (for instance, as a senator, a mayor, a member of a committee,

a president of a society or of the university), or in passing (for instance, a student sitting an examination), or in being the winner (in a contest or a lottery, at Bingo, in a sports competition, in a game of cards or chess, or anything else) and so on? Will a vote turn out to be favourable – for instance, for a given law, or for an issue of confidence facing a government and so on? Is the accused in a given trial really the murderer? And, in any case, will he be convicted as such? Is the approaching tram the one I am waiting for? Will the next child of a given couple be a boy? Will it rain tomorrow at a given place? Will the next attempt at a soft landing on the moon be successful?

In all cases, and in various ways, if we want to be more detailed, or to extend the questions, we often conveniently express ourselves in terms of random quantities. In the examples of elections and voting, we might ask the following sorts of questions. How many votes are favourable? How many against, invalid or abstentions? What is the percentage of those in favour? In the case of examinations, contests and competitions, what is the mark or position obtained? And when – what year, day or moment – will the event in question occur (moon launch, trial verdict, vote, birth of the particular baby etc.)? Or, alternatively, how many will succeed – among those participating in an examination, contest, sports event and so on? Which one among them – identified by entry number, or position in alphabetical order – will attain first place, or second place? Who, within a given age limit, will be best placed? In a competition with several stages, or legs, who among the entrants will maintain, or improve, or worsen, their position with respect to the previous placings?

In other cases one uses different terminology. Random point: for instance, the point of the lunar surface which will next be reached. Random set: the set of those who pass an examination, the set of points on the earth's surface on which rain will fall tomorrow, the set of instants at which the temperature at a given place is below, above, or at, zero. Random function: the temperature at the above-mentioned place, the score during a competition, the number of votes of confidence since a certain date and so on, all considered as functions of time. If one wishes to avoid reference to irrelevant items of information (for instance, by referring to an entry number rather than to the individual concerned) it is preferable to speak of a multi-event, rather than a random quantity, and so on.

2.10.3. It is clear that in all cases it would be possible to go into more and more detail, and if all the cases we have mentioned were considered simultaneously we would arrive at even more minute subdivisions. And to these cases could be added others, ad infinitum. To arrive at a final subdivision into 'points' – not further divisible – would at least imply the construction of all possible 'histories of the universe', distinct in every detail. These would include, for example, the precise specification, instant by instant, of the position of every atom, and of the thoughts and moods of each individual – including, possibly, beings, more or less similar, living on other worlds. Even if we limit ourselves to much more restricted problems, an exhaustive description, though very much reduced in scale, would by no means turn out to be more realistic. Consider a single toss of a coin: unimaginable faculties would be needed if we wished to provide a description, with such absolute precision, of a single one of the possible ways in which a person tosses the coin, the air influences the movement, and every peculiarity of the ground and of the coin at the point and position of the latter's fall gives rise to successive movements, and so on, until the coin comes to rest. But this would still be nothing, because, instead, we must imagine and distinguish the totality of such ways.

We have pushed ourselves to absurd lengths – in a way pointless in itself – but perhaps this will serve to illustrate the thesis that it is inappropriate to distinguish between events represented by 'points', or by 'sets', thinking of it as something systematic, rather than being dependent on momentary conveniences of representation.

2.10.4. This has been said to emphasize the considerations already made (in Section 2.7.7 and elsewhere), but it is even more necessary to underline the sense in which an event (random quantity etc.) has to be – as we said – something 'well determined'. This means that the formulation must be unambiguous and complete, in such a way as to rule out any possibility of argument (for instance in the case of a bet which is based on it). To give an example: 'A.N. Other wins the lottery'[28] is an event only if the person A.N. Other, of whom we are speaking, is perfectly individuated, along with the circumstances that make the statement precise. Examples of the latter might be: win in next week's drawing; or in the first week that he plays; or any week of this year; and so on. It should also be made precise, or understood, whether possible wins in partnership with others are to be included, or not, along with any other possible aspects allowing ambiguity. By changing the individual, or any of the circumstances or provisos, we obtain other events, all different from each other. We say this only to avoid the situation where, being familiar with other terminologies, someone might think that they should be called 'identical events' or, even worse, 'trials' of 'the same event', which consists in 'winning the lottery'.

Conversely, two events expressed in completely different ways are identical – that is they are the same event – if we know that the occurrence of either one of them implies the occurrence of the other. Suppose, for instance, that we know for certain that this week A.N. Other is going to play the 'straight' three numbers 21–63–82 on the Roman wheel, and nothing else: in this case, the two events 'A.N. Other is going to win the lottery this week' and 'This week the numbers 21, 63 and 82 will come out on the Roman wheel' are identical. On the other hand, in order to demonstrate that it would be wrong to think in terms of the identification of a 'fact', we note the following: 'A.N. Other is going to win the next time he plays' and 'Next week the youngest person playing is going to win' are two *distinct events* which might, by chance, turn out to be the *same fact* if next week A.N. Other plays and wins and, in addition, happens to be the youngest player. This example also serves the purpose of making clear that there is no need to identify explicitly the person and the drawing (either by the date or, possibly, the wheel) so long as, by some means or other, it turns out that whether we must call the *statement* true or false is well determined.

2.10.5. One could object, with reason, that such a requirement is practically unrealizable, and that, in fact, it is not even realized in the example which we have just given. For instance, how is the statement 'A.N. Other is going to win the lottery the next time he plays' to be evaluated if A.N. Other never plays again for the rest of his life? This should be made clear by means of some arbitrary convention. In most cases of this kind, however, we shall interpret the statement in a sense which falls outside the present concept of an event, but which leads to a generalization (conditional event) that we will consider

28 *Translators' note*. Every Saturday in Italy, at each of ten cities, a drawing takes place of 5 from 90 possible numbers. To enter the lottery, one places a bet, prior to the drawing, specifying which combination(s) of numbers (up to a maximum of 5) one thinks will be drawn in a chosen city. The device which produces the numbers is known as a 'wheel'.

explicitly later on (in Chapter 4). In addition to being *true* (=1) or *false* (=0) it could also be *void* (= Ø). In terms of a bet, this means that it could not only result in gain or loss, but could also, in certain cases, be called off. If these things are not made clear explicitly, in a systematic way, then even the statement 'A.N. Other is going to win the lottery next week' might appear ambiguous, because of the doubt as to whether we mean 'false' or 'void' if A.N. Other does not play: in such a case one implicitly assumes certain refinements, but without justification. We will not labour this point, postponing further discussion until the appropriate place. In the same way, we do not enter into discussion of certain other questions, like perhaps the preceding ones, which may appear sterile but which, if misunderstood, give rise to numerous possible ambiguities and errors. In contrast to the above, these questions can be put off until later. Let us merely remark – in order not to seem mysterious – that, above all, it is a question of discussing the actual possibility of obtaining, within a given time and with greater or lesser certainty and precision, information concerning the events and quantities of interest, about which we are at present uncertain.

2.10.6. We now return to the examples that we considered before in order to draw attention to some of the kinds of problems that we will frequently meet in the future, and which will serve, for the time being, to illustrate the notions introduced in the preceding paragraph.

When we ask how many of the participants in an examination will succeed in passing, we have an example of a problem concerning the *number of successes*, $Y = E_1 + E_2 + \ldots + E_n$, where E_h = 'the success of participant h' or, alternatively, one concerning the frequency, or percentage, of successes, Y/n. Other examples, chosen from the infinite number of possibilities, might include the following: the number of 'white balls in n given drawings from an urn'; or of 'males among the first n births registered in Orvieto next year'; or of 'those among the n participants in a competition with many stages who maintain, after a given stage, their previous position.'

Clearly, Y can only assume the values 0, 1, 2,…, n, and, obviously, these will all be actually possible if the events E_h are logically independent. This means that the set of all those who pass an examination can, in fact, be any one of the 2^n subsets of candidates (including the whole set and the empty set); that is for each h = 0, 1, 2,…, n, all the $\binom{n}{h}$ subsets of h individuals are subsets for which $Y = h$. In the cases of examinations, drawings from an urn, births and so on, this will be true under most of the usual assumptions (and we shall see what these are shortly, when we turn to counterexamples). For the time being, however, we note that the $n + 1$ values can all be possible, even in cases where logical independence does not hold. Suppose, for instance, that E_h means that 'the person placed in the hth position in a competition has reached some minimum prescribed score' (or time in a race, distance with a throw, height with a jump). It is possible that all, or none, or any intermediate number h, will succeed; in the latter case these are obviously the first h and no others. We do not have logical independence since if E_h is true, all the preceding ones are necessarily true, and if false all the following ones are false.

At the other extreme, it is possible that Y is certain. This is the case, for instance, if E_1, E_2,…, E_n represent the drawing of white balls in n successive drawings *without replacement* from n balls, h of which are white; then we certainly have $Y = h$, the number h being known with certainty at the present moment. But in every case (drawings with replacement, examinations, sex of births) we find ourselves in the same situation if we are acquainted with the outcome as a whole, even though ignorant of the results of

single drawings and so on. It is important to notice that the E_h are, in this case, not only logically, but also *linearly*, *dependent* ($E_1 + E_2 + \ldots + E_n = Y = h$). The logical dependence assumes a concrete form in the fact that once all the white balls (or all the others) are out, the result of the subsequent drawings is certain (in any case, this is always so for the last drawing at least).

All intermediate hypotheses can be shown to be possible by the use of examples of a more or less artificial nature. The actual possibility of all $n + 1$ values is also compatible with linear dependence: if $n \geqslant 3$ we could have $E_1 = E_2$ with certainty if one thinks, for example, of the first two balls being drawn from an urn containing balls of the same colour in pairs. Restrictions on Y may exist in the case of competitive examinations with a maximum number of awards available, or in the case of drawings without replacement of n balls from an urn containing N balls of which H are white : in this case $n - (N - H) \leqslant Y \leqslant H$, the restrictions being real if the limits are >0 and $<n$, respectively.

An important case, of some interest since it is rather less obvious, is that in which all values except $n - 1$ are possible. We meet it in the example of 'maintaining rank in a classification', which, more abstractly, consists in considering the elements that remain fixed under a permutation. One of the many well-known different interpretations is the following: we put, more or less haphazardly, n letters into n envelopes, and we consider the random quantity Y which denotes the number of letters correctly placed. Clearly, all outcomes are possible, except that of making just one error: one letter cannot be misplaced if all the others are in their own envelopes, since only the correct envelope then remains.

2.10.7. In the case of three or more alternatives (for each of n multi-events[29]) we must consider for each of them the number of successes or realizations: for instance, X, Y, Z, with $X + Y + Z = n$, X, Y, Z being the number of votes for, against or abstentions, out of n votes; or wins, draws and losses out of n games; or of bachelors, married men or widowers out of n males; and so on, and so forth. Similarly for cases involving many alternatives: for example $X_1 + X_2 + \ldots + X_6 = n$ for occurrences of the points 1, 2,..., 6 when we throw n dice, or a single die n times (or, in the previous example, if we distinguish marital status and sex).

Problems of this kind are called problems of *subdivisions*: here we have been dealing with the subdivisions of the integer n into a given number of (non-negative) integer summands, but more generally we could consider subdivisions of a given quantity q into any kinds of summands whatsoever – non-negative real values $X_1 + X_2 + \ldots + X_m = q$. We often prefer to take $q = 1$, that is to reduce to percentages: in the preceding case we could also divide the numbers of occurrences by n, obtaining in this way the frequencies. A classical example is the subdivision of an interval (into m parts with $m - 1$ division points). One could also imagine, however, the masses of the m parts into which an object of mass q breaks on falling; or, alternatively, the masses of m materials from which it is constructed (for example m metals if we are dealing with an alloy). We shall meet these kinds of problems again.

It is of interest to note that in such cases the m random quantities are linearly dependent. Other quantities that have to be considered in connection with questions of this nature

29 Of course, this is also valid in the case of only two alternatives; in this case, however, it is trivial to take into account the number of occurrences of each of them since $Y = n - X$.

are also linearly dependent if they are linear combinations of them. As examples, we note the difference between votes for and against, or the total number of 'points' scored (taking 2 for a win, 1 for a draw). On the other hand, this would not be true, for example, for ratios, such as votes in favour divided by votes against, where one would have logical but not linear dependence.

2.10.8. In the above example of a ratio ($Z = Y/X$), and in others that will follow, the logical dependence will be functional dependence (in the clearest case, with $f(X_1, X_2,\ldots, X_n)$ such that each X_h turns out to be uniquely determined within the permitted field). Naturally, the given definition does not imply anything of this kind. Not only may the uniqueness fail – as when we consider points on the spherical surface $X^2 + Y^2 + Z^2 = 1$ with admissible values not constrained to be non-negative – but one might also consider all points of the sphere as possible (by substituting $\leqslant$ for =) without destroying the logical dependence. To see this, note that, given $X = x$ and $Y = y$, the possible values of Z lie in the segment between $\pm\sqrt{(1 - x^2 - y^2)}$, which is a function of x and y. Given that X, Y, Z can all assume values between ± 1, we have logical independence only for the case in which all the points of the cube $-1 \leqslant x, y, z \leqslant +1$ are possible: the exclusion of a single point, for example the origin, is sufficient to give logical dependence (to avoid it, we would have to exclude the points on the coordinate planes; i.e. the value 0 for each random quantity separately). One also has logical dependence if one excludes from the cube the points for which, for example, $X + Y + Z$ (or XYZ, XY/Z, etc.) is rational, or transcendental, or whatever (to avoid it, one should instead exclude separately, X, for example, being rational, Y being transcendental, Z being zero).

2.10.9. A case of logical dependence, which is of practical importance and frequent occurrence, is the following: given a number of random quantities, say X, Y, Z, we denote, by definition, the smallest of these by X, the middle one by Y, the greatest by Z. In this case, we exclude all those points which are not included in the dihedron $y - x \geqslant 0$, $z - y \geqslant 0$, even if the coordinates of the points are possible values for X, Y, Z (unless all the possible values for X are less than all the possible values for Y, and these are less than all the possible values for Z, in which case $X \leqslant Y \leqslant Z$ does not constitute a restriction). It is necessary to pay attention to circumstances of this kind, as the necessity of establishing and taking appropriate account of them could be overlooked.

If we take the example of a subdivision resulting from the splitting of a fallen object – let us say into three pieces, X, Y, Z – the situation differs according to whether the criterion by which we rank them is the order of magnitude, or something else not depending on it. For instance, we might take the angle formed between the half-line starting from the point of fall and passing through the barycentre of the piece in question and the direction North, the angle being taken in a counterclockwise direction.

The same thing holds in the example we are about to consider now, where X, Y, Z are the sides of a random prism (rectangle): for example, a block of stone, a building, a suitcase. We may or may not have more or less 'natural' circumstances which lead us to define, in each case, what we mean by 'length' (X), 'breadth' (Y) and 'height' (Z). Without getting bogged down in an analysis, which everyone can provide for themselves anyway, the answer seems to be easy for the suitcase, not always such for the building – the distinction between length and breadth may not be clear if there is no recognizable façade – and indeterminate for the block (unless we use conventions

based on how it is temporarily situated with respect to North, East and the zenith). If we agree to call the maximum side the length, and the minimum side the height, we are in the other situation.

Given this random prism – and however we think of the problem, with the sides X, Y, Z logically independent or not – let us consider its diagonal U, area V and volume W. In either case, these are random quantities that are logically (and even, in a unique way, functionally) dependent on the preceding ones: $U = \sqrt{(X^2 + Y^2 + Z^2)}$, $V = 2(XY + XZ + YZ)$, $W = XYZ$. Clearly, however, the dependence is not *linear*; when we return to the question, in Chapter 3, this example will serve to clarify, in an appropriate way, how, and why, certain reasonings about uncertainty, though seemingly obvious, are correct in some cases, but not in others (and this according to whether one has linear dependence or not).

2.11 Concerning Certain Conventions of Notation

2.11.1. As we announced in Section 2.5.3, and briefly mentioned in Chapter 1 (1.9.3 and 1.9.4), we will demonstrate, by means of examples, the utility that can be derived in many cases from the use of conventions introduced in the present chapter for simplifying the notation. To be explicit:

- the identification of TRUE and FALSE with 1 and 0;
- the 'lattice' operations for numbers.

2.11.2. The convention TRUE = 1 and FALSE = 0 turns out to be very useful also when applied outside of the field of events, to propositions or any 'conditions' whatsoever

Examples. $(x \geqslant a)$ is the function which = 0 for $x < a$ and = 1 for $x \geqslant a$; we could write such a function as $F(x) = (x \geqslant a)$, and, more generally,

$$F(x) = \sum_h p_h \left(x \geqslant a_h\right)$$

is the step-function with jumps p_h at the points $x = a_h$; assuming that the a_h are in increasing order of magnitude, this could also be written

$$F(x) = \sum_h c_h \left(a_h \leqslant x < a_{h+1}\right),$$

which denotes that in the given interval the value is

$$c_h = \sum_i p_i \left(i \leqslant h\right) = \sum_{i=1}^{h} p_i .$$ [30]

In the last example we used the function $(a \leqslant x < b)$, which is = 1 in the given interval and = 0 outside: more generally, we use $(x \in I)$ to denote the indicator function of the set I (the function which = 1 if x is in I, and = 0 otherwise).

---

30 Given the purely illustrative purpose of these forms of notation, we omit all the possible refinements that should be added, case by case, in specific applications: for instance, here, hypotheses of convergence if we are dealing with series: in an opposite sense the convention $a_{n+1} = \infty$ if a_n is the last term, etc. The notation $\leqslant$ instead of $<$ etc. will vary from case to case.

Using such a function as a multiplier, one obtains immediately the restriction of a function to a given interval or set; for example,

$$x^2\,(x \geq 0) = 0 \text{ for } x \leq 0 \text{ and } = \mathrm{x}^2 \ \textit{for}\ x \geq 0;$$

$f(x) = x(1 - x)(-1 \leqslant x \leqslant 1) = x(\mathrm{l} - x)(|x| \leqslant 1)$ is equal to $x(\mathrm{l} - x)$ for x in $[-1, 1]$, 0 otherwise; and, more generally, for a function with a different expression in different intervals, for example,

$$f(x) = a(x+3)^3\,(-3 \leqslant x < -1) + (b - cx^2)(-1 \leqslant x < 1) + a(3-x)^3\,(1 \leqslant x < 3),$$

or even (for a large, or infinite, number of intervals)

$$f(x) = \sum_h f_h(x)(h \leqslant x < h+1), \text{ or } f(x) = \sum_h f_h(x)(a_h \leqslant x < a_{h+1}).$$

Remarks. The examples in which the functions are denoted by $F(x)$ and $f(x)$, respectively, can be interpreted as the *distribution* function (F) and the *density* function ($f = F'$) of a distribution. These notions may already be familiar but will, in any case, be introduced in Chapter 6.

2.11.3. In the previous cases of summation, we have already seen the expression of the condition functioning as a multiplier in order to define each single sum-function, or (under another equivalent interpretation) specifying, for a given x, which terms had to be summed. The systematic usage of such a convention to this end, even in the absence of a useful interpretation in the first sense, would seem to be very convenient, both for clarity and typographical convenience. It replaces, in an advantageous manner, either explanations in the text, or complicated instructions to be composed under the summation (or integral) sign and so on.

The meaning of the following examples is self-evident:

$$\sum a_h(h \in H),\ \sum a_h(b_h \in B),\ \sum a_h(h \neq 0),\ \sum a_{hk}(h \neq k),\ \sum a_{hk}(h \leqslant k),$$

$$\int f(x)\mathrm{dx}(2n \leqslant x \leqslant 2n+1), \qquad \int f(x,y)\mathrm{dxdy}(x^2 + y^2 \leqslant r^2).$$

2.11.4. *Use of the Boolean operations*. The Boolean operations $\vee$ and $\wedge$ often serve (even better than the system given above) to denote '*truncations*' and similar operations. For instance, the function $F(x)$ = '*x provided it is not less than zero or greater than one*' could be written in either of the two ways

$$F(x) = x(0 \leqslant x \leqslant 1) + (x > 1) = 0 \vee x \wedge 1,$$

and the second is clearly simpler. In general, the function which = $f(x)$ but is never less than m or greater than M can be written as $m \vee f(x) \wedge M$, and similarly $m(x) \vee f(x) \wedge$

$M(x)$, so long as we always have $m(x) < M(x)$ (otherwise we would not have $(m \vee f) \wedge M = m \vee (f \wedge M)$, and the notation would not be admissible).

This notation, in our context, will serve in particular for random quantities: we present here a few examples in both notations (and the Boolean form seems to be simpler):

$$
\begin{aligned}
X(X \geqslant 0) &= 0 \vee X, \qquad X(X \leqslant 0) = 0 \wedge X, \\
X &= X(X \geqslant 0) + X(X \leqslant 0) = 0 \vee X + 0 \wedge X, \\
|X| &= X(X \geqslant 0) - X(X \leqslant 0) = 0 \vee X - 0 \wedge X; \\
X(0 \leqslant X \leqslant K) + K(X > K) &= 0 \vee X \wedge K, \\
X(|X| \leqslant K) + K\big[(X > K) - (X < -K)\big] &= -K \vee X \wedge K,
\end{aligned}
\Big\} \; (K > 0)
$$

and so on.

3

Prevision and Probability

3.1 From Uncertainty to Prevision

3.1.1. So far, even in the way we presented the preceding examples, we have limited ourselves to depicting and representing the situation facing You, when You are interested in distinguishing among a more or less extensive class of alternatives (all those which, in the present state of your information, appear possible to You). This preliminary topic, which we will have to consider more deeply in what follows, is still within the ambit of ordinary logic, the logic of certainty. One should always be careful to distinguish clearly between those things belonging to this domain and those belonging to the *probabilistic* domain – the ambit of the *logic of uncertainty*, the logic of *prevision* – to which we must now turn our attention. It was precisely in order to pin-point this distinction that we decided upon this form of exposition, presenting concepts and related examples which reveal the situation as it is, while leaving undetermined all questions concerning the possible introduction of probability, its conceptual basis and its evaluation. It would certainly be easier, and seemingly more instructive, to go right ahead and take the two steps together, instead of just the one. In other words, we could present right away, fused together in the examples and definitions, both the probability (which answers the need) and the uncertainty (from which the need arises), without first making such a need 'felt', and then pausing to reflect upon it. It is precisely this latter course, however, which must be recommended.

The situation is this: having distinguished the possible cases, and having represented them in the way which seems to You most effective (or in any way convenient to You), if You then wish to restrict yourself to the logic of certainty You have to stop, and consider the question closed. Is this what You *want* to do? And *can* You do it?

For each one of us, it is often the case that we *do not content ourselves* (or are not *able* to content ourselves) with this, and therefore we proceed further. And, strictly speaking, to proceed further means to enter into what we have called the logic of *prevision* (in a sense that we will make clear in order to draw attention to the distinction between this and other interpretations, whose drawbacks must be pointed out).

3.1.2. *Prevision, not prediction*. In order to use this word, 'prevision', it will be necessary to give an absolutely precise meaning to it (and to derived words) and to insist on this meaning and keep it in mind, consistently and scrupulously, in the sequel. It must be distinguished, and in fact contraposed, to another word, which, in everyday language,

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

is perhaps more commonly attributed to it, and for which we will reserve the alternative name, '*prediction*'.

To make a *prediction* would mean (using the term in the sense we propose) to venture to try to 'guess', among the possible alternatives, the one that will occur. This is an attempt often made, not only by would-be magicians and prophets, but also by experts and such like who are inclined to precast the future in the forge of their fantasies.[1] To make a 'prediction', therefore, would not entail leaving the domain of the logic of certainty, but simply including the statements and data which we assume ourselves capable of guessing, along with the ascertained truths and the collected data. It is not enough to tone down the 'prophetic' character of such pronouncements by taking precautions with feelings ('I think', 'perhaps', etc.) as we have already mentioned: either these artificial additions remain without any authentic meaning, or they need to be actually translated into probabilistic terms, substituting prevision in place of prediction.

If we remain within the logic of certainty, such additions not only have no authentic meaning in themselves, but, in point of fact, they render meaningless the entire discussion. If the discussion affirms that something is *true*, and the 'perhaps' means that instead of being true it could also be *false*, this is equivalent to retracting the preceding statement, declaring it to be invalid and unfounded (cancelling it, disowning it). If not, then 'perhaps' should be erased as it might give a false impression of such a retraction.

Alternatively – and this is the approach indicated below, and which corresponds to the subjectivistic conception of probability – the 'perhaps' can be explained as an indication, even if crudely qualitative, of a degree of subjective probability which, if we wished, could be made more precise, and even quantified.

All this would be very clear if there did not exist, unfortunately, in the very field of probability and statistics, certain tendencies to avoid the choice, playing precisely on that ambiguity which we drew attention to earlier, and making it worse. In fact, the ambiguity of the 'perhaps' (which could be innocent, due to simple unwariness) is fraudulently concealed beneath a showy exterior. It is translated into technical terms like 'accept' and 'reject', which neither mean YES or NO with certainty, nor are to be interpreted in a probabilistic sense, but simply lay claim to be themselves 'accepted', rather than 'rejected', without giving to those terms any 'acceptable' meaning whatsoever.

3.1.3. *Prevision*, in the sense in which we have said we want to use this word, does not involve guessing anything. It does not assert – as prediction does – something that might turn out to be true or false, by transforming (over-optimistically) the uncertainty into a

---

1 Everyone will no doubt have noticed, and had occasion to notice, how often the 'foresights of experts' turn out to be completely different from the facts, sometimes spectacularly so. In the main, this is precisely because they are intended as predictions which 'deduce', more or less logically, a long chain of consequences – still considered necessarily plausible – from the assumed plausibility of an initial hypothesis. Interesting examples of the lack of connection between prevision and reality (in the political field) are pointed out and discussed by B. de Jouvenel in 'Futuribles', *Bulletin Sedeis*, 20 January 1962.

Here also one might note the irrelevant distinction, as far as prevision and prediction are concerned, between the future and the past: the hypothetical reconstructions of murders or historical facts made by detectives, scholars or novelists, based on scanty data and meriting varying degrees of respect, are, in the above sense, 'predictions'.

It is useful to ask oneself, incidentally, whether such 'facile fantasies' are really 'rich fantasies', or rather 'poor fantasies', in that ineptitude or laziness prevents us from seeing how many other possibilities there are, besides the first one we happened to think of.

claimed, but worthless, certainty. It acknowledges (as should be obvious) that what is uncertain is uncertain: in so far as statements are concerned, all that can be said beyond what is said by the logic of certainty is illegitimate. If we think that something might be added, if we think, as we remarked above, that we can *proceed further*, it will necessarily be a question of entering into a completely new field and scheme of things, one which goes beyond the logic of certainty, even if it must be linked to it and superimposed upon it.

When we cease to content ourselves with the logic of certainty, in what sense do we go beyond it? In what sense do You go beyond it? Let us ask ourselves this question. Ask yourself. The thing we are not content with, and neither are You, is the agnostic and undifferentiated attitude towards all those things which, not being known to us with certainty, are uncertain, are possible. There are no degrees[2] of possibility: it is possible (equall possible) that it snows on a winter or summer day; that a great champion or a novice wins the competition; that every student, whether well-prepared or not, will pass an examination; that next Christmas You will find yourself at any place in the world. However, You do not content yourself with this, and, in fact, it is not your real attitude. Faced with uncertainty, one feels, and You feel too, a more or less strong propensity to expect that certain alternatives rather than others will turn out to be true; to think that the answer to a certain question is YES rather than NO; to estimate that the unknown value of a certain quantity is small rather than large.

These attitudes, of ours and of yours, do not lead us – as in the case of someone who claims to make a spot-on prediction – to assert as certain or impossible something which, on the basis of the logic of certainty, is possible but uncertain, and which remains such whatever further assertions or thoughts might be added. Uncertain things remain uncertain, but we attribute to the various uncertain events a greater or lesser degree of that new factor which is extralogical, subjective and personal (mine, yours, his, anybody's), and which expresses these attitudes. In everyday language this is called *probability*, a concept that we shall have to clarify and study. *Prevision*, in the sense we give to the term and approve of (judging it to be something serious, well-founded and necessary, in contrast to prediction), consists in considering, after careful reflection, all the possible alternatives, in order to distribute among them, in the way which will appear most appropriate, one's own expectations, one's own sensations of probability.

We all of us enter into this ambit of prevision in a spontaneous fashion; sometimes without a specific need, for the sole reason that one is interested in the object of uncertainty, that there are desires or hopes that certain alternatives occur, anxieties and fears regarding the occurrence of unfavourable alternatives, and that the weighing up of such hopes and fears matters to one. Sometimes, on the other hand, all one's behaviour may necessarily depend on a comparative evaluation, albeit crude and perhaps unconscious, of the various impending risks, and of the various targets that one can set oneself. In this sense, and because of the enormous range of possibilities, one may find oneself *compelled* to weigh up such evaluations, and to express the prevision. In the case of more important and conscious decisions, one might try to *reason about* each choice, and *weigh up the pros and cons* by means of some criterion or other.

---

2 In a certain sense, however, there exists a partial order since one could call, 'not less possible' than another, an event which is a consequence of it (in the same way as one could call, 'not less extended' than another one, a set containing it). In both cases, however, no step forward is made towards a comparability or measurability of the 'possibilities' or of the 'extensions'.

3.1.4. *Coherence*. It is precisely in investigating the connection that must hold between evaluations of probability and decision making under conditions of uncertainty that one can arrive at criteria for measuring probabilities, for establishing the conditions which they must satisfy, and for understanding the way in which one can, and indeed one must, '*reason about them*'. It turns out, in fact, that there exist simple (and, in the last analysis, obvious) conditions, which we term conditions *of coherence*: any transgression of these results in decisions whose consequences are manifestly undesirable (leading to certain loss).

The 'one must' is to be understood as 'one must if one wishes to avoid these particular objective consequences'. It is not to be taken as an obligation that someone means to impose from the outside, nor as an assertion that our evaluations are always automatically coherent. On the contrary, it is precisely because this is an area where it is particularly easy to slip into incoherence that it is important to learn the *art of prevision* (to adapt the phrase *Ars Conjectandi*, used by James Bernoulli as the title of the first treatise on the calculus of probability).

Given any set of events whatsoever, the conditions of coherence impose no limits on the probabilities that an individual may assign, except that they must not be in contradiction amongst themselves. Without further delay, we will proceed to the construction of the theory of probability, using as a basis the theory of decision making. For the time being, this will be done in an extremely simplified form, as a preliminary clarification of ideas. In the next paragraph we will discuss certain other aspects of the problem, and then turn to the constructive formulation.

Within this framework, we obtain the greatest insight by considering as a starting point the case of random quantities (especially when we interpret them as random gains). With a more rigorous approach, inspired by decision-theoretic considerations, it is essentially a question of returning to that problem of a *fair evaluation*, or estimation, which, in connection with similar problems of an economic nature, seems to have foreshadowed by centuries the beginnings of the calculus of probability. In this sense, the modern setting of the problem, within decision theory, constitutes, to some extent, a return to its origins.

The definition of the probability of an event will turn out to be contained automatically in that given in the case of random quantities: we simply define events as particular random quantities. From a mathematical viewpoint also, this would appear the appropriate thing to do. In Chapter 2 we saw that in the case of events the most useful arguments, which are very simple if one considers the events as special points in the space of random quantities, are not available if one thinks in terms of the set of events without reference to the space in which this set *is* 'naturally' embedded, and in which it is necessary to *see it* embedded.

Let us proceed then to the matter in hand, starting with the consideration of a *random gain X*: by this we mean a random quantity X having the meaning of gain (the latter intended, of course, in an algebraic sense; a loss is a negative gain). The possible values of X could, therefore, also be negative, either in part, or entirely. We might ask an individual, for example You, to specify the *certain gain* that is considered *equivalent* to X. This we might call the *price* (for You) of X (we denote it by $\mathbf{P}(X)$[3]) in the sense that, on

3 We could write $\mathbf{P}_i(X)$ to emphasize that we are dealing with the evaluation of a particular individual i. This is an unnecessary precaution, however, since it is understood that we are always referring ourselves to the evaluation of a given individual (real or fictitious).

your scale of preference, the random gain X is, or is not, preferred to a certain gain x according to whether x is less than or greater than $\mathbf{P}(X)$. For every individual, in any given situation, the possibility of inserting the degree of preferability of a random gain into the scale of the certain gains is obviously a prerequisite condition of all decision-making criteria. Among the decisions that lead to different random gains, the choice must be the one that leads to the random gain with the highest price. Moreover, this is not a question of a condition but simply of a definition, since the price is defined only in terms of the very preference that it means to measure, and which must manifest itself in one way or another.

In general, it is not true that if one is prepared to buy an article A at the price $\mathbf{P}(A)$ and an article $\boldsymbol{B}$ at the price $\mathbf{P}(B)$, one must be prepared to buy both of them together at a price $\mathbf{P}(A) + \mathbf{P}(B)$. It may happen that the purchase of one of them affects, in various ways, the desirability of the other. Similar qualifications hold if instead of two articles A and B we consider two random gains X and Y; this case will be examined in the next paragraph. In both cases, however, additivity is something more than just an interesting simplifying hypothesis, which may be approximately valid. As we shall see later, provided we modify slightly the way in which the notion of price $\mathbf{P}(X)$ is introduced, additivity will turn out to be an exact property, the foundation of the whole treatment.

3.1.5. *Properties of* **P**. If You are indifferent to the exchange of X for $\mathbf{P}(X)$ and of Y for $\mathbf{P}(Y)$, then, if we assume the simplifying hypothesis given above, You are also indifferent to the exchange of $X + Y$ for $\mathbf{P}(X) + \mathbf{P}(Y)$. The value for which You are indifferent to the exchange of $X + Y$ is, however, by definition, $\mathbf{P}(X + Y)$; we therefore conclude that

a) *the price* **P** *is an additive function*:

$$\mathbf{P}\big(X+Y\big)=\mathbf{P}\big(X\big)+\mathbf{P}\big(Y\big). \tag{3.1}$$

A second property, obvious, but equally fundamental, can be derived by noting that $\mathbf{P}(X)$ must not be less than the lower bound of the set of possible values for X, $\inf X$, nor greater than the upper bound, $\sup X$ (otherwise the choice would allow a certain loss). Therefore,

b) *the price* **P** *must satisfy the inequality:*

$$\inf X \le \mathbf{P}\big(X\big) \le \sup X; \tag{3.2}$$

obviously, this condition only imposes a restriction if the random quantity X is *bounded* in at least one direction (either $\inf X > -\infty$ or $\sup X < +\infty$). Generally, but not always, we will restrict our attention to the bounded case (i.e. bounded from above and below).

When we come to formulate and examine this set-up in a more exhaustive fashion, we shall see that the two extremely simple conditions, (*a*) and (*b*), are *not only necessary but also sufficient* for coherence – that is for avoiding undesirable decisions. This is all that is needed for the foundation of the whole theory of probability: in fact, the definition of probability immediately reduces, as a special case, to that of a price **P**.

We observe, from (*a*) and (*b*), that the price **P** must also be a *linear* function, in the sense that for every real a we have

$$\mathbf{P}(aX) = a\mathbf{P}(X),^{4} \tag{3.3}$$

and therefore, more generally,

$$\mathbf{P}(aX + bY + cZ + \ldots) = a\mathbf{P}(X) + b\mathbf{P}(Y) + c\mathbf{P}(Z) + \ldots \tag{3.4}$$

for any *finite* number of summands.

Given this property, it is possible to extend the definition of $\mathbf{P}(X)$ to the case in which X is a random quantity (pure number), or a random magnitude not having the meaning of gain (for instance, time, length, etc.). In fact, it suffices to choose a coefficient a whose dimension is such that aX is a monetary value: for instance, in the cases of time and length we could take Lire/s and \$/cm. We now define $\mathbf{P}(X) = (1/a)\mathbf{P}(aX)$: this is well defined, since the expression is invariant with respect to the choice of a (we can substitute λa in place of a, where λ is a nonzero real number).

In the general case (where we do not have a monetary value), the term 'price' is no longer appropriate: we speak instead of the '*prevision of X*', valid in all cases,[5] and, in particular, of the '*probability of E*' when $X = E$ is an event.

The probability $\mathbf{P}(E)$ that You attribute to an event E is, therefore, the certain gain p that You judge equivalent to a unit gain conditional on the occurrence of E: in order to express it in a dimensionally correct way, it is preferable to take pS equivalent to S conditional on E, where S is any amount whatsoever, one Lira or one million, \$20 or £75. Since the possible values for a possible event E satisfy $\inf E = 0$ and $\sup E = 1$, for such an event we have $0 \leq \mathbf{P}(E) \leq 1$, while necessarily $\mathbf{P}(E) = 0$ for the impossible event, and $\mathbf{P}(E) = 1$ for the certain event.[6]

3.2 Digressions on Decisions and Utilities

3.2.1. In Section 3.1, we have introduced the notions of *prevision* and *probability* by following the path laid down by certain decision-theoretic criteria of an essentially economic nature: the presentation was, however, in a simplified form.

It follows, therefore, that before going any further we should make some comments and give some further details about the theory of decision making, and above all about *utility*. The latter, together with probability, is one of the two notions on which the correct criterion of decision making depends. We warn the reader, however, that this is in

4 This is obvious if a is rational, and the extension to every a is straightforward if X is always positive (because then if a lies between a' and a'', we also have aX between $a'X$ and $a''X$). But we can always write $X = Y - Z$, where $Y = X$ ($X \geq 0$) and $Z = -X$($X \leq 0$), and these numbers are always non-negative: $Y = X$ if $X > 0$ and zero otherwise, $Z = -X$ if $X < 0$ and zero otherwise. The conclusion is therefore valid for Y and for Z, and hence for $X = Y - Z$.

5 This corresponds to 'mathematical expectation' in classical terminology, and to 'mean value' in more up-to-date usage. We prefer to reserve the term 'mean value' for *objective* distributions (e.g. statistical distributions).

6 These are the only cases in which the evaluation of the probability is predetermined, rather than permitting the choice of any value in the interval from 0 to 1 (*end-points included*). The predetermination that one meets in these cases arises because there exists no uncertainty and the use of the term probability is redundant. The same thing holds for prevision: $\mathbf{P}(X)$ necessarily has a given value x if and only if X has x as a unique possible value; i.e. if X is not really random. The above is the special case where either $x = 0$ or $x = 1$.

the nature of a digression and anyone not interested in the topic can skip it without any great loss: the details (of a noneconomic nature) that are given in Section 3.3, and in subsequent sections, will prove quite sufficient.

3.2.2. *Operational definitions*. In order to give an effective meaning to a notion – and not merely an appearance of such in a metaphysical–verbalistic sense – an operational definition is required. By this we mean a definition based on a criterion that allows us to measure it.[7] We will, therefore, be concerned with giving an operational definition to the prevision of a random quantity, and hence to the probability of an event.

The criterion, the operative part of the definition which enables us to measure it, consists in this case of testing, through the *decisions* of an individual (which are observable), his *opinions* (previsions, probabilities), which are not directly observable.

Every measurement procedure and device should be used with caution, and its results carefully scrutinized. This is true in physics, despite the degree of perfection attainable, and even more so in a field as delicate as ours, where similar and much more profound difficulties are encountered.

In the first place, if, as is implicit in what we have said so far, we identify, generically, decisions and preferences, then we are ignoring many of the extraneous factors that play a part in decision making. Nobody accepts all the opportunities or bets that he judges favourable, and perhaps we all sometimes enter into situations that we judge unfavourable. To reduce the influence of such factors it is convenient to effect the observations on the phenomena isolated in their most simple forms: this is in fact what we attempt to do when we construct measuring devices. For the purpose of a formal treatment of the topic, we will present (in the next section) two different procedures by means of which we try to force the individual to make conscious choices, releasing him from inertia, preserving him from whim. Of course, we have to establish that the two procedures are equivalent, and this we shall do.

A doubt might remain, however. Are the conclusions that we draw after observing the actual behaviour of an individual, directly making decisions in which he has a real interest, more reliable than those based on the preferences which he expresses when confronted with a hypothetical situation or decision? Both the direct interest and the lack of it might on the one hand favour, and on the other obstruct, the calmness and accuracy, and hence the reliability, of the evaluations. In any case, it is not really a mathematical question: it is useful to be aware of the problem, but it is mainly up to the psychologists to delve further into the matter. We merely note that between the two extreme hypotheses one could consider an intermediate one that might be of interest; the case of an individual being consulted about a decision in which others are interested. This might well lead to responsibility in the judgment without affecting the calmness of the decision maker. In Chapter 5, 5.5.6, we encounter another example which is similar in spirit to the last one: this is where the accuracy of the evaluation is related to one's self-respect in some competitive situation (with prizes which are materially insignificant, but which are related to the significance of the competition).

At this point, the reader may be wondering on what basis individuals do evaluate their probabilities or previsions: the question is not appropriate, however. Firstly, we must

7 See P. Bridgman, *The Logic of Modern Physics*, Macmillan, New York (1927).

attempt to discover opinions and to establish whether or not they are coherent. Only at the second stage, having acquired the necessary knowledge, could we also apply it to investigate these other aspects, and not until it was very much advanced could this be done in a sufficiently satisfactory way (up to and including the rather complicated justifications for the case of evaluations based on frequencies, a case wrongly considered simple).

3.2.3. *Reservations concerning rigidity*. The main question that we have to face in these 'remarks' is the one already mentioned when we expressed reservations about assuming additivity for the price of a random gain: recall that it is this hypothesis which underlies the definition of prevision, and the special case of probability.

It is well known, and indeed obvious, that usually this is not realistic because of the phenomenon of risk aversion (or occasionally its opposite, but we shall not bother with such cases). In fact, as we already noted in effect when we introduced it, the hypothesis of additivity expresses an assumption of *rigidity* in the face of risk. Let us now try to make this clear. As a preliminary, it will suffice to restrict ourselves to simple examples that are within our present scope. These will be sufficient to show that in order to obtain a formulation which is completely satisfactory from the economic point of view, it is necessary to eliminate such rigidity by introducing the notion of *utility*. On the other hand, they will also show that one is able to manage without this notion, except when occupied with applications of an expressly economic nature.

Suppose that You are faced with two eventualities that You judge equally probable: taking the standard example, it could be a question of Heads or Tails. Given the hypothesis of rigidity in the face of risk, You should be indifferent between 'receiving with certainty a sum S, or twice the sum if a particular one of the two possible cases occurs': likewise, between 'losing with certainty a sum S, or twice the sum if a particular one of the two possible cases occurs'; and similarly between 'accepting or not accepting a bet which, in the two possible cases, would lead either to a loss, or to a gain, of the same sum S' This much is obvious, but in any case we shall carry out the calculations as an exercise. Let us denote by A and B the two events: $A + B = 1$ because one and only one of the two occurs. Their probabilities, being supposed equal, must each have the value $\frac{1}{2}$, since $\mathbf{P}(A) = \mathbf{P}(B)$, and $\mathbf{P}(A) + \mathbf{P}(B) = \mathbf{P}(A + B) = \mathbf{P}(1) = 1$. It follows that cases of so-called indifference simply imply the equality of the following: S and $(2S)/2$, $-S$ and $-(2S)/2$, 0 and $\frac{1}{2}S + \frac{1}{2}(-S)$ (since, for instance, the gain $2S$ conditional on the event A is the random quantity $X = 2SA$, $\mathbf{P}(X) = \mathbf{P}(2SA) = 2S\mathbf{P}(A) = 2S.\left(\frac{1}{2}\right)$.

If instead, as is likely, You are risk averse, then in all cases You will prefer the *certain* alternative to the *uncertain* one (the form and extent of the aversion will depend upon your temperament, or perhaps be influenced by your current mood, or by some other circumstance). To arrive at the actual indifference, You would content yourself with receiving with certainty a sum S' (less than S) in exchange for the hypothetical gain $2S$; You would be disposed to pay with certainty a sum S'' (greater than S) in order to avoid the risk of a hypothetical loss $2S$; You would pay a certain penalty K in order to be released from any bet where the gain and loss are, in monetary terms, symmetric.

This means, however, that by virtue of risk aversion one has symmetry in the scale in which one's judgments of indifference are based: that is equal levels in passing from 0 to S' and from S' to $2S$, or in passing from $-2S$ to $-S''$ and from $-S''$ to 0, or in passing from $-S$ to $-K$ and from $-K$ to S. The scale no longer coincides with the monetary one, as in the

case of rigidity. In short, as far as we are concerned, things proceed as if successive increments of equal monetary value had for You smaller and smaller subjective value or *utility*. This term – often used in a similar sense, but in a questionable form, in economic science – has been rehabilitated and adopted with the specific meaning derived from the present considerations about risk.

3.2.4. *The scale of utility*. The above considerations enable us to construct a scale of utility; that is a function $U(x)$, the utility of the gain x, whose increments, $U(x_{i+1}) - U(x_i)$, are equal when, and only when, we are indifferent between the corresponding increments of monetary gain, $x_{i+1} - x_i$. We could proceed, for instance, by dividing an interval into two 'indifferent increments', in the way indicated in the examples above, and in the same way obtain subdivisions into 4, 8, 16,..., parts. It would be more appropriate, instead of considering the variable x representing the gain, to take $f + x$, where f is the individual's 'fortune' (in order to avoid splitting hairs, inappropriate in this context, one could think of the value of his estate). Anyway, it would be convenient to choose a less arbitrary origin in order to take into account the possibility that judgments may alter because in the meantime variations have occurred in one's fortune, or risks have been taken, and in order not to preclude for oneself the possibility of taking these things into account, should the need arise. Indeed, as a recognition of the fact that the situation will always involve risks, it would be more appropriate to denote the fortune itself by F (considering it as a random quantity), instead of with f (a definite given value).

What we have said concerning the scale of utility makes it intuitively clear – and this is sufficient for the time being – that if, in order to define 'price', we refer to this scale rather than to the monetary scale, then additivity holds. In fact, one might say that such a scale is by definition the monetary scale deformed in such a way as to compensate for the distortions of the case of rigidity which are caused by risk aversion. The formulation put forward in the preceding paragraph could, therefore, be made watertight, and this we will do shortly by working in terms of the utility instead of with the monetary value. This would undoubtedly be the best course from the theoretical point of view, because one would construct, in an integrated fashion, a theory of decision making (of the criteria of coherent decisions, under conditions of certainty or uncertainty), whose meaning would be unexceptionable from an economic viewpoint, and which would establish simultaneously and in parallel the properties of *probability* and *utility* on which it depends. The fundamental result lies, in fact, in recognizing that *the criteria of coherent decision making are precisely those which consist of the choice of any evaluation of the probabilities and any utility function* (with the necessary properties) *and in fixing as one's goal the maximization of the prevision of the utility*. Of course, it is possible to behave coherently with respect to decisions and preferences without knowing anything about probability and utility. The fact is, however, that in this case one must behave *as if* one is acting in the above manner, *as if* obeying an evaluation of probability and a scale of utility underlying one's way of thinking and acting (even if without realizing it). Provided one could succeed in exploring these activities in an appropriate way, it might be possible to trace back and individuate the two components.

This unified approach to an integrated formulation of decision theory in its two components was put forward by F.P. Ramsey (1926) and rigorously developed by L.J. Savage (1954). However, there are also reasons for preferring the opposite approach,

the one which we attempt here. This consists in setting aside, until it is expressly required, the notion of utility, in order to develop in a more manageable way the study of probability.

3.2.5. *An alternative approach*. The idea underlying this alternative approach stems from the observation that the hypothesis of rigidity, as considered above, is acceptable in practice – even if we stick to monetary values – provided the amounts in question are not 'too large'. Of course, the proviso has a relative and approximate meaning: relative to You, to your fortune and temperament (in precise terms, to the degree of convexity of your utility function U); approximate because, in effect, we are substituting in place of the segment of the curve U which is of interest the tangent at the starting point. Clearly, the smaller the range considered, the more satisfactory is the approximation. With this in mind, we might consider replacing the previous definition of $\mathbf{P}(X)$ – which we temporarily distinguish, denoting it by $\mathbf{P}^*(X)$ – with a new one, which we define by means of the relationship:

$$\mathbf{P}(X) = \lim_{a \to 0} \left(\frac{1}{a} \right) \mathbf{P}^*(aX).$$

Instead, we prefer a less orthodox but more natural and manageable solution, which consists of not changing anything, but merely remarking that in economic examples one must remain within appropriate limits (which, as an aid to understanding, we call 'everyday affairs').

There are several reasons behind this choice (and, more generally, behind rejecting the standard method of considering both probability and utility together, right from the very beginning).

Firstly, on a purely formal level, there is an objection to taking the passage to the limit so seriously as to base a definition on it: in fact, if a becomes too small an evaluation loses, in practice, any reliability. This is the same phenomenon that one encounters when attempting to define density, although the underlying reasons are different. One needs to consider the ratios mass/volume for neighbourhoods small enough to avoid macroscopic inhomogeneities, but not so small as to be affected by discontinuities in the structure of matter. We accept that once we are in the area of mathematical idealization we can leave out of consideration adherence to reality in every tiny detail: on the other hand, it seems rather too unrigorous to act in this way when formulating that very definition that should provide the connection with reality.

This does not mean that it is not useful to accept the form of the passage to the limit (as an innocuous and convenient assumption, although not appropriate to fulfil the function of a definition). In any case, let us suppose that we have introduced the linear prevision $\mathbf{P}(X)$, and that we know the utility function U, which, for the sake of simplicity, we now take to be expressed as a function of the gain x. Then the original $\mathbf{P}(X)$ as it actually turns out to be, assuming that the hypothesis of rigidity is not satisfied (this is denoted above by $\mathbf{P}^*$, but from now on we denote it by $\mathbf{P}_U$), can be expressed immediately as a transform of $\mathbf{P}$ by means of U:

$$\mathbf{P}_U(X) = U^{-1}\left\{\mathbf{P}\left[U(X)\right]\right\}. \tag{3.5}$$

In the standard case, where U is convex, we have $\mathbf{P}_U(X) < \mathbf{P}(X)$, as noted above, and as can be seen from the theory of associative means (Chapter 2, 2.9.3). In order to be able to distinguish between the two concepts, when referring to them, we will say that:

a transaction is depending on whether	$\mathbf{P}_U$	*indifferent*, remains constant,	*advantageous*, or increases,	*disadvantageous* or decreases;
a transaction is depending on whether	$\mathbf{P}$	*fair*, remains constant,	*favourable*, or increases,	*unfavourable* or decreases.

A fair transaction is such for everyone agreeing on the same evaluation of probabilities, even for the other contracting party ($\mathbf{P}(-X) = -\mathbf{P}(X)$); an indifferent transaction is not such as U varies, and cannot be such for both contracting parties if they both have convex utility functions (in this case $\mathbf{P}_U(-X) < \mathbf{P}(-X) = -\mathbf{P}(X) < -\mathbf{P}_U(X)$).

3.2.6. *Some further remarks*. Finally, let us turn to the other reasons for preferring this approach: these are essentially concerned with simplicity. The separation of probability from utility, of that which is independent of risk aversion from that which is not, has first of all the same kind of advantages as result from treating geometry apart from mechanics, and the mechanics of so-called rigid bodies without taking elasticity into account (instead of starting with a unified system).

The main motivation lies in being able to refer in a natural way to combinations of bets, or any other economic transactions, understood in terms of monetary value (which is invariant). If we referred ourselves to the scale of utility, a transaction leading to a gain of amount S if the event E occurs would instead appear as a variety of different transactions, depending on the outcome of other random transactions. These, in fact, cause variations in one's fortune, and therefore in the increment of utility resulting from the possible additional gain S: conversely, suppose that in order to avoid this one tried to consider bets, or economic transactions, expressed, let us say, in 'utiles' (units of utility, definable as the increment between two fixed situations). In this case, it would be practically impossible to proceed with transactions, because the real magnitudes in which they have to be expressed (monetary sums or quantity of goods, etc.) would have to be adjusted to the continuous and complex variations in a unit of measure that nobody would be able to observe.

Essentially, our assumption amounts to accepting as practically valid the hypothesis of rigidity with respect to risk: in other words, the identity of monetary value and utility[8] within the limits of 'everyday affairs'. One should be concerned, however, to check whether this assumption is sufficiently realistic within a wide enough range: actually, it seems safe to say that under the heading of 'everyday affairs' one can consider all those transactions whose outcome has no relevant effect on the fortune of an individual (or firm, etc.), in the sense that it does not give rise to substantial improvements in the situation, nor to losses of a serious nature.

There is no point in prolonging this discussion, but it seems appropriate to mention an analogy from economics, and one from insurance: these – in the same spirit as the

8 Except for (obviously inessential) changes of origin and unit of measurement.

preceding geometrical–mechanical analogy – are sufficient to clarify the question, both from a conceptual and practical point of view. Using the prices $\mathbf{P}(X)$ as they appear in our hypothesis of rigidity is to do the same thing as one does in economics when one considers the total price of a set of goods, of given amounts, on the basis of the unit prices in force at the time, without taking into account the variation that a possible transaction would cause by changing the supply and demand situation. On the other hand, these variations are only noticeable if the quantities under consideration are sufficiently large. Even more apposite is the example of actuarial mathematics: indeed, the latter is nothing other than a special case of the theory we are discussing. In the main, it is traditionally concerned with the terms of an insurance under *fair* conditions ('pure' is the usual terminology: pure premium, etc.), and only in special cases – for instance, the theories dealing with the risk of the insurer, or with the advantage for those exposed to risk in insuring themselves – does one speak in terms of utility (or something equivalent, if such a notion is not introduced explicitly). Notwithstanding the fact that this stems less from deliberate choice than from a tradition that lacks an awareness of the questions involved, the 'rigid' approximation has turned out to be satisfactory for the greater part of this most classical field of application of the calculus of probability to economic questions. We intend to use only the simplified version; the above considerations suggest that this is a reasonable thing to do.[9]

On the other hand, we shall see (in Chapter 5) how, although starting from the hypothesis of rigidity, one can arrive at the evaluation of probabilities by means of criteria which are neutral with respect to it. The method, which takes as its basis the most meaningful concept, and then clarifies it by means of this simplifying hypothesis, therefore achieves its objective without prejudicing the conclusions.

3.3 Basic Definitions and Criteria

3.3.1. We must now translate into actual definitions and proofs those things that we have hitherto put forward in an introductory form, bringing in any necessary refinements, and beginning the developments.

We have given some idea of what a *prevision function* **P** is, and what conditions it must satisfy in order to be *coherent*. The function **P** represents the opinion of an individual who is faced with a situation of uncertainty. To each random magnitude X, there corresponds the individual's evaluation $\mathbf{P}(X)$, the *prevision* of X, whose meaning, operationally, reduces, in terms of gain, to that of the (fair) *price* of X. This includes, in particular, the special case of *probability* (which is the more specific name given to prevision when X is an event). A prevision function **P** is *coherent* if its use cannot lead to an inadmissible decision (i.e. such that a different possible decision would have certainly led to better results, whatever happened). We have remarked already that coherence reduces to linearity and convexity.

3.3.2. In order to fix the formulation in a precise way, we will now put forward two *criteria* (in the sense of *devices* or *instruments* for obtaining a measurement). Each one furnishes an *operational definition* of probability or prevision **P**, and,

9 For all these topics see de Finetti–Emanuelli (1967), Part I.

together with the corresponding *conditions of coherence*, can be taken as a foundation for the entire theory of probability.

Let us recall that the term 'operational' applies to those definitions that allow us to reduce a concept not merely to sentences, which might have only an apparent meaning, but to actual experiences, which are at least conceptually possible. Think of Einstein's definition of 'simultaneity' by means of signals: until that time no-one had even doubted that the term lacked an absolute meaning. That definitions should be operational is one of the fundamental needs of science, which has to work with notions of ascertained validity, in a pragmatic sense, and which must not run the risk of taking as concepts illusory combinations of words of a metaphysical character.

In our case, for the definition of $\mathbf{P}(X)$, it is a question of stating exactly what 'the rules of the game' are. To state, in other words, what, in the application of a given criterion, are the practical consequences that You know You must accept, and which You do accept, when You enunciate your evaluation of $\mathbf{P}(X)$ (whose meaning as 'price' is already essentially given). From a conceptual point of view, in the case of coherence too the pointers given in Section 3.1.5 are sufficient in themselves. To make them explicit in a compact form for specific criteria provides, however, a more incisive schematization of the theory by reducing it to a really small nucleus of initial assumptions.

3.3.3. As far as the extension of the domain of definition of a function of prevision **P** is concerned, we assume that in principle **P** could be evaluated (by You, by anybody) for *every* event E or random quantity X: this is in contrast to what is assumed in other theories and so it is appropriate to point it out explicitly. It will be sufficient for You to place yourself under the restriction of a certain criterion, which we shall soon make explicit, and being forced to answer – that is to make a choice among the alternatives at your disposal – to reveal your evaluation of $\mathbf{P}(X)$ (or, in particular, of $\mathbf{P}(E)$). This is valid, as we have said, *in principle*: in other words, we intend not to acknowledge any distinction according to which it would make sense to speak of probability for some events, but not for others.[10] On the other hand, however, we certainly do not pretend that **P** could actually be imagined as determined, by any individual, for *all* events (among which those mentioned or thought of during the whole existence of the human race only constitute an infinitesimal fraction, even though an immense number). On the contrary, we can at each moment, and in every case, assume or suppose **P** as defined or known for all (and only) the random quantities (or, in particular, events) belonging[11] to some completely arbitrary set $\mathscr{X}$: for instance, those for which we know the evaluation explicitly expressed by the individual under consideration.

Without leaving this set, whatever it may be, we can recognize whether or not **P** includes any incoherence; if so, the individual, when made aware of this fact, should eliminate it, modifying his evaluations after reconsideration. The evaluation is then coherent and can be extended to any larger set whatsoever: the extension will be uniquely determined up to the point that the coherence demands and is, to a large extent, more or less arbitrary outside that range. One can only proceed, therefore, by interrogating the individual and alerting him if he violates coherence with respect to the preceding evaluations.

---

10 For instance, the two following distinctions are quite common: yes for 'repeatable' events, no for 'single' instances; yes if X belongs to a measurable set I, no otherwise. See Appendix.

11 And not, necessarily, belonging to something reducible to a ring (or to a σ-ring) of events. Again, see Appendix.

Among the answers that do not make sense, and cannot be admitted, are the following: 'I do not know', 'I am ignorant of what the probability is', 'in my opinion the probability does not exist'. Probability (or prevision) is not something that in itself can be known or not known: it exists in that it serves to express, in a precise fashion, for each individual, his choice in his given state of ignorance. To imagine a greater degree of ignorance that would justify the refusal to answer would be rather like thinking that in a statistical survey it makes sense to indicate, in addition to those whose sex is unknown, those for whom one does not even know 'whether the sex is unknown or not'.

Other considerations and restrictions may enter in if we consider functions of probability defined other than as an expression of the opinion of a given individual. If, for instance, after having considered and interrogated many individuals, we want to study 'their common opinion', **P**, this will only exist in the domain of those X for which all the $\mathbf{P}_i(X)$ coincide (in this way defining $\mathbf{P}(X)$), and will not exist elsewhere. We can also confine ourselves (there is nothing to prevent us) to evaluations which conform to more restrictive criteria to which one would prefer to limit the investigation, excluding in this way events for which one would like to say that the probability 'does not exist' or 'is not known', knowing all along that such motivations remain, nonetheless, meaningless within the present formulation. I may please a friend of mine by not inviting along with him a person whom he judges 'a jinx', without myself believing that such things exist, nor understanding how others can believe in it.

As far as *coherence* is concerned, we will again underline here in what sense the notion is and must be *objective*. The conditions of coherence must exclude the possibility of certain consequences whose unacceptability appears expressible and recognizable to everyone, independently of any opinions or judgments they may have regarding greater or lesser 'reasonableness' in the opinions of others. Let this be said in order to make clear that such conditions, although *normative*, are not (as some critics seem to think) unjustified impositions of a criterion that their promoters consider 'reasonable': they merely assert that 'you must avoid this if you do not want …' (and there follows the specification of something which is obviously undesirable). We will see this immediately – note it well! – in the two criteria we are about to put forward.

3.3.4. *Criteria for the evaluations*. We now present the details of the two criteria mentioned above; each will consist of the following:

- a *scheme of decisions* to which an individual (it could be You) can subject himself in order to reveal – in an operational manner – that value which, *by definition*, will be called his *prevision* of X, or in particular his *probability* of E,
- and a *condition of coherence* that enables one to distinguish (so that the distinction has an objective meaning) whether an individual's set of previsions is coherent, and therefore acceptable, or, conversely, intrinsically contradictory.

In both cases, the prevision of X will be a value $\overline{x}$, which can be chosen at will as an 'estimate' of X; along with such a choice goes the necessity of making precise, according to which scheme is used, the otherwise completely indeterminate[12] meaning of the

12 Or, even worse, open to being interpreted as 'prediction'!

word 'estimate'. To anticipate the outcome in words, both criteria start by considering the random magnitude given by the difference, or deviation, $X - \bar{x}$, between the actual value X and that chosen by You. Both lead to the same $\bar{x}$ if applied coherently.

The first criterion stipulates that You must accept a bet proportional to $X - \bar{x}$, in whatever sense chosen by your opponent (i.e. positively proportional either to $X - \bar{x}$ or to $\bar{x} - X$). This means that there is no advantage to You in deviating, one way or the other, from the value that makes the two bets indifferent for You; otherwise, one or other would be unfavourable to You, and the opponent could profit from this by an appropriate choice.[13]

The second criterion stipulates that You will suffer a penalty (positively) proportional to the square of the deviation $(X - \bar{x})^2$, increasing as one deviates in either direction from the actual value.

This is evident if one recalls the properties of the barycentre (stable equilibrium, minimum of the moment of inertia), which give an analogy and, in fact, a perfect interpretation. Those who already know something about probability or statistics will be well acquainted with the fact that these properties characterize $\mathbf{P}(X)$. The latter is usually called '(mathematical) expectation', or 'mean value', and is denoted by $\mathbf{E}(X)$ or $\mathbf{M}(X)$: the only novelty lies in making use of it as an operational and direct definition of $\mathbf{P}(X)$, and in particular of probability. Given the probabilities of all possible values (if they are finite in number), it is clear how $\mathbf{P}(X)$ can be expressed as a function of them: the extension of this result to the general case will be immediate when we introduce the notion of a 'probability distribution' (see Chapter 6). In the latter approach, however, one introduces the simpler notion (that of 'prevision') by means of the more complicated one (that of 'distribution'), which itself becomes a prerequisite, and forces us to use more advanced mathematical tools (Stieltjes integrals) than necessary. The same thing happens in the case of a solid body: the barycentre is easily determined and, it might be said, is always useful; the exact distribution of mass can never be determined in practice, and is of relatively little interest.

Two further remarks. Firstly, let us recall 'the hypothesis of rigidity with respect to risk', which we continue to assume in what follows (not without noting where appropriate, under the heading of 'Remarks', any implications of this hypothesis at those points where it merits attention). In order to fulfil more easily the resulting requirement of considering only 'moderate amounts', and to omit certain delicate points which are better reconsidered later on (in Section 3.11, etc.), we restrict ourselves for the time being to *bounded* random magnitudes (i.e. those whose possible values are contained in some interval; in other words, $-\infty < \inf X$, $\sup X < +\infty$).

Concerning preferences for one or other of the two criteria of definition, it is merely a question of individual taste, since – as we have stated, and will later show – the two definitions (together with their respective conditions of coherence) are equivalent. The first has a meaning that is slightly more immediately intuitive, but, as far as actual deductions are concerned, the second is more meaningful and fits better into a decision-theoretic framework. A third criterion, which has useful applications, will be derived in Chapter 5, but does not lend itself to an autonomous presentation.

13 This is the same criterion as 'divide the cake into two parts and I will choose the larger', which ensures that the person dividing it does so into parts he judges to be equal.

3.3.5. *The first criterion.* Given a random quantity (or random magnitude) X, You are obliged to choose a value $\bar{x}$, on the understanding that, after making this choice, You are committed to accepting any bet whatsoever with gain $c(X - \bar{x})$, where c is arbitrary (positive or negative) and at the choice of an opponent.

Definition. $\mathbf{P}(X)$, the prevision of X according to your opinion, is by definition the value $\bar{x}$ that You would choose for this purpose.

Coherence. It is assumed that You do not wish to lay down bets which will with *certainty* result in a loss for You.[14] A set of your previsions is, therefore, said to be *coherent* if among the combinations of bets which You have committed yourself to accepting there are none for which the gains are *all uniformly negative*.[15]

Analytic conditions. Expressed mathematically, this means that we must choose the values $\bar{x}_i = P(X_i)$ such that there is no linear combination

$$Y = c_1\left(X_1 - \bar{x}_1\right) + c_2\left(X_2 - \bar{x}_2\right) + \ldots + c_n\left(X_n - \bar{x}_n\right)$$

with sup Y negative (conversely, inf Y cannot be positive, because then sup $(-Y) = -\inf Y$ would be negative).

Remark. Observe the objective character of these conditions, revealed by the fact that only 'possible values' are referred to.

3.3.6. *The second criterion.* You suffer a penalty L[16] proportional to the square of the difference (or deviation) between X and a value $\bar{x}$, which You are free to choose for this purpose as you please:

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

(where k, arbitrary, is fixed in advance, possibly differing from case to case).[17]

Definition. $\mathbf{P}(X)$, the prevision of X according to your opinion, is the value $\bar{x}$ which You would choose for this purpose.

Coherence. It is assumed that You do not have a preference for a given penalty if You have the option of another one which is *certainly* smaller. Your set of previsions is therefore said to be *coherent* if there is no other possible choice which would certainly lead to a uniform reduction in your penalty.

Analytic conditions. The definition of coherence implies that there exist no values x_i^* which, when substituted for the chosen $\bar{x}_i = \mathbf{P}(X_i)$, lead to the penalty

---

14 Giving rise to what is sometimes called a 'Dutch Book'.

15 The reason why we cannot simply say 'all negative' (i.e. <0), but must add 'uniformly' (i.e. $< -\varepsilon$ with ε positive) will be given later (for the time being we do not worry about the finer points). By 'combinations' we always mean linear combinations *of a finite number* of the bets even if there are infinitely many of them).

16 From *Loss*, the terminology introduced by A. Wald.

17 It is convenient to think of k as being homogeneous with X, so that the expression turns out to be a pure number; with the further understanding that we multiply by a monetary unit u, L has the dimension of a monetary value. This avoids the complication of writing it, or assuming it as included in k, by conjuring up a strange factor of dimension $u^{1/2}$.

$$L^* = \sum_i \left(\frac{X_i - x_i^*}{k_i} \right)^2$$

being uniformly less than

$$L = \sum_i \left(\frac{X_i - \bar{x}_i}{k_i} \right)^2,$$

for *any possible points* $(X_i, X_2, \ldots, X_n)$; that is belonging to the set $\mathcal{D}$.

Remark. As for the first criterion.

3.3.7. *The equivalence of the two criteria*. The identity of the previsions given by the two criteria can be verified immediately.

Let $\bar{x}$ be the prevision of X based on the first criterion, and $\bar{\bar{x}}$ that based on the second; this implies, respectively, that:

i) in the first case, the random gain X is judged equivalent to the certain gain $\bar{x}$ (hence: preferable to each $x < \bar{x}$, but not to any $x > \bar{x}$);
ii) in the second case, the gain $-(X - \bar{\bar{x}})^2$ – negative, since a penalty ! – is judged preferable to any other $-(X - x)^2$ with $x \neq \bar{x}$; in other words, the gain

$$G = (X - x)^2 - (X - \bar{\bar{x}})^2$$

is preferred to 0 (for all $x \neq \bar{\bar{x}}$).

More generally, let us compare preferences between the penalties corresponding to any two values of x, say $x = a$ and $x = b$, and let us denote by $c = \frac{1}{2}(a + b)$ the mid-point of the interval $[a, b]$.

The choice of a is preferred to that of b, if the gain $G = (X - b)^2 - (X - a)^2$ if preferred to 0; in other words, expanding, if

$$G = (X^2 - 2bX + b^2) - (X^2 - 2aX + a^2) = 2(a - b)X - (a^2 - b^2)$$
$$= 2(a - b)(X - c)$$

is preferred to 0. Preferring G to 0 means that $\mathbf{P}(G) > 0$; on the basis of the first criterion it turns out that $\mathbf{P}(G) = 2(a - b)(\bar{x} - c)$, an expression which is positive if $a > b$ and $\bar{x} > c$, or, conversely, if $a < b$ and $\bar{x} < c$. In other words, in either case, if $\bar{x}$ lies in the subinterval between c and a; that is if $\bar{x}$ is *closer* to a than it is to b.

Our assertion is an obvious corollary of this result (which it seemed useful to put forward in this more general form): the optimal choice, $x = \bar{\bar{x}}$, is given by $\bar{\bar{x}} = \bar{x}$.

The equivalence of the conditions of coherence can also be verified by expansions of this sort (and we shall do so, writing them out in full, for those who would like to check them and apply them directly). Conceptually, however, we can make everything incomparably easier, and intuitively meaningful, by presenting an obvious geometrical interpretation.

3.4 A Geometric Interpretation: The Set $\mathscr{P}$ of Coherent Previsions

3.4.1. Any prevision in the linear ambit $\mathscr{A}$ of the n random quantities $X_1, X_2,\ldots, X_n$ consists in fixing, in the n-dimensional space with coordinates $x_1, x_2,\ldots, x_n$ (the linear ambit $\mathscr{A}$), the n values $\overline{x}_1,\overline{x}_2,\ldots,\overline{x}_n$, where $\overline{x}_i = \mathbf{P}(X_i)$, and hence corresponds to a point in the said space. The conditions of coherence state – as we shall immediately verify – that *the set $\mathscr{P}$ of coherent previsions is the closed convex hull of the set $\mathscr{Q}$ of possible points*.

For the first criterion: in a form that is more directly suited to the purpose in hand, the necessary and sufficient condition for coherence can be expressed by saying that *every linear relation* (*or inequality*) *between the random quantities* X_i

$$c_1X_1 + c_2X_2 + \ldots + c_nX_n = c \ \left(\text{or } \geqslant c\right)$$

must be satisfied by the corresponding previsions $\mathbf{P}(X_i)$:

$$c_1\mathbf{P}\left(X_1\right) + c_2\mathbf{P}\left(X_2\right) + \ldots + c_n\mathbf{P}\left(X_n\right) = c \ \left(\text{or } \geqslant c\right).$$

Geometrically, a point P represents a coherent prevision if and only if *there exists no hyperplane separating it from the set $\mathscr{Q}$ of possible points*; this characterizes the points of the convex hull.

For the second criterion: here one introduces into the (*affine*!) linear ambit $\mathscr{A}$ a *metric* of the form $\rho^2 = \sum_i \left(x_i/k_i\right)^2$, setting

$$\begin{aligned}\text{'penalty'} = L &= \left(P - Q\right)^2 \\ &= \text{'the square of the distance between the prevision} \\ &\quad \text{-point } P \text{ and the outcome-point } Q, \text{ according to the given metric'}.\end{aligned}$$

The necessary and sufficient condition for coherence requires, in geometrical terms, that *P cannot be moved in such a way as to reduce the distance from all points Q*; this is another characterization of the convex hull.[18] Further explanations, and diagrams in the simple cases, are given in Chapter 5, Section 5.4.

3.4.2. *Other interpretations*. Every prevision-point P, which is admissible in terms of coherence, is a barycentre of possible points Q_j with suitable weights (or is a

---

18 If we move the point P to another position P^*, its distance from a generic point A increases or decreases depending on whether A is on the same side as P or P^*, with respect to the hyperplane that bisects the segment PP^* orthogonally.

If P is not in the convex hull of $\mathscr{Q}$ there exists a hyperplane separating it from $\mathscr{Q}$. Moving P to P^*, its orthogonal projection onto such a hyperplane, diminishes its distance from all points Q in $\mathscr{Q}$ (which are on the opposite side). More precisely, the diminution of the penalty, $L - L^*$, i.e. the square of the distance, is always at least $(P - P^*)^2$: in fact, $(Q - P)^2 - (Q - P^*)^2 = [(Q - P^*) - (P - P^*)]^2 - (Q - P^*)^2 = (P - P^*)^2 - 2(Q - P^*) \times (P - P^*)$, and this scalar product is negative, since the component of the first vector parallel to the second is in the opposite direction.

Suppose instead that P belongs to the convex hull of $\mathscr{Q}$. Then to whatever point P^* we move P, it always follows that for some point Q the distance increases: if, with respect to the bisecting hyperplane, they were all on the same side as P^*, the point P would be separated from the convex hull of $\mathscr{Q}$, contrary to the hypothesis.

limit-case[19]). On the other hand, the possible points are themselves particular cases of previsions; degenerate cases, in that the probability is concentrated at a unique point Q_j. In words, one could say, according to this interpretation, that *a prevision turns out to be a mixture of possibilities*.

Of course, one can also form linear combinations of different coherent previsions (with non-negative weights, summing to 1) again obtaining coherent previsions. More generally, if $\mathscr{P}_0$ is any set of coherent previsions, then its closed convex hull is also a set of coherent previsions, the *mixtures* of those in $\mathscr{P}_0$. Let us denote it by $\mathscr{P}_1$.

3.5 Extensions of Notation

It is convenient, in addition to being natural (and also useful for compactness of notation), to exploit the linear structure of **P** in order to extend the range of applicability of this symbol to any random elements whatsoever belonging to a linear space (vectors, matrices, *n*-tuples of numbers or magnitudes, functions, etc.), or even just to a linear manifold (a linear subspace which also contains the zero: for example, the points of a space in which the differences between points $\mathbf{u} = A - B$, constitute a linear space of vectors).

As a formal definition, it is sufficient to state that **P** is always intended to be *linear*, so that if f is any linear function – that is $f(A)$ is a scalar linear function of the points or elements A of our linear space or manifold – we have $f(\mathbf{P}(A)) = \mathbf{P}(f(A))$. For practical purposes, it is enough to note that **P** operates on the components or coordinates, so that, if

$$A = 0 + X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}\big(or, \text{ in conventional notation, } A = \big(X, Y, Z\big)\big),$$

we could write

$$\mathbf{P}\big(A\big) = \mathbf{P}\big(X\big)\mathbf{i} + \mathbf{P}\big(Y\big)\mathbf{j} + \mathbf{P}\big(Z\big)\mathbf{k} = \overline{x}\mathbf{i} + \overline{y}\mathbf{j} + \overline{z}\,\mathbf{k}$$

(in other words, $\mathbf{P}(X, Y, Z) = (\mathbf{P}(X), \mathbf{P}(Y), \mathbf{P}(Z))$).

A case of particular importance is the following: if Z is a *complex* random quantity, and we denote by X and Y, respectively, the real and imaginary components, its prevision will be

$$\mathbf{P}\big(Z\big) = \mathbf{P}\big(X + \mathrm{i}Y\big) = \mathbf{P}\big(X\big) + \mathrm{i}\mathbf{P}\big(Y\big) = \overline{x} + \mathrm{i}\overline{y}.$$

As a practical rule, it is sufficient to replace the random component X by the corresponding prevision $\overline{x}$; for example:

$$\mathbf{P}\big(X_1, X_2, \ldots, X_n\big) = \big(\overline{x}_1, \overline{x}_2 \ldots, \overline{x}_n\big), \qquad \mathbf{P}\big(\|X_{rs}\|\big) = \|\overline{x}_{rs}\|, \quad \text{etc.}$$

---

19 To be precise: either they can be obtained as barycentres of most $n + 1$ points Q_j (in the n-dimensional space), or they are *adherent* points of $\mathscr{Q}$ (but not belonging to $\mathscr{Q}$). For instance, we could think of $\mathscr{Q}$ as the set of points on the circumference of a circle, having rational angular distance from one of its points (with respect to the complete angle). We are in the plane, $n = 2$, and each point inside the circle is inside triangles with vertices in $\mathscr{Q}$: hence it is the barycentre of $3 = n + 1$ points (two would suffice if it were on chords connecting rational points, and only one if it coincided with such a point). The points, which are on the circumference, but not rational, are required in order to complete the closed convex hull: they are adherent points of $\mathscr{Q}$ (i.e. there are points of $\mathscr{Q}$ in each of their neighbourhoods).

In the case of a random function $X(t)$ (where t is the independent variable; for example, time) it would appear to be an unnecessary subtlety (but it is not) to say that one could write $f = \mathbf{P}(X)$ to mean that f is the function which for each t gives $f(t) = \mathbf{P}(X(t))$. It would be a little more explicit to write $f(\cdot) = \mathbf{P}(X(\cdot))$ in order to indicate that it is a question of operating on a variable whose position is denoted by the point. Here, however (at least if one does not want to be limited to considering not more than a finite number of t_h at once), one would step outside of the ambit in which, for the time being, we have expressed our intention of remaining.

3.6 Remarks and Examples

3.6.1. The properties that we have established in Section 3.4 could be said to contain the whole calculus of probability, even though we have not as yet mentioned probability, except to point out that it is a special case of prevision. Sections 3.8 and 3.9 will be devoted to this special case, giving it the attention it merits. However, from our point of view it turns out to be better formulated and much clearer if embedded in the general case, where the basic properties present themselves as simple, clear and 'practical'. It is precisely for this reason (and certainly not because of any dubious motive of wishing to start, come what may, by showing off, with no justification, the greatest generality and abstraction) that we did not begin the discussion with the case of events, and have still not stopped to consider it. Otherwise, we would have found ourselves, at this moment, having defined $\mathbf{P}(E)$ and not $\mathbf{P}(X)$, in more difficulty than if we had defined a unique concept, and with the unavoidable problem of producing $\mathbf{P}(X)$ as something not equally immediate, but as the combination of the $\mathbf{P}(E)$ and who knows what mathematical definition of integral.

3.6.2. *Some remarks concerning the two criteria*. Every operational definition, if one wants to take it too seriously as an actual method of measurement, carries with it the difficulty that the discussions of principle become mixed up with the doubts deriving from the practical imperfections inherent in any tool or procedure (these, however, often arise for reasons which may be important). Let us accept that this difficulty is unavoidable but that it is by no means a tragedy, since the definition deals with an idealized case, or limit-case, of conceptually possible experiments. Having said this, and having repeated that it is always infinitely better than any attempt at a mere verbal definition, emptily 'philosophical', there remains, nonetheless, the necessity of making oneself aware of the weak points in order to keep in mind the appropriate precautions.

We have already discussed, in Section 3.2, 'rigidity in the face of risk'; in other words, the temporary identification of utility and monetary value. At this juncture, a brief mention, with specific reference to the two criteria that we have put forward, will suffice. They both assume, implicitly, the hypothesis of rigidity. In the first place, they take the different bets, which are used as 'tests', to be summable; to be rigorous, the stipulation of any one of them should modify slightly the conditions for the stipulations of the others; secondly, by virtue of having a homogeneous character, in the sense that the procedure itself presupposes that $\mathbf{P}(aX) = a\mathbf{P}(X)$ (this adds no further restrictions, it is the same rigidity). This is useful if one attempts to limit the bets to be of moderate size; it is dangerous to allow it to be used indiscriminately. In the first criterion one has to think

that in practice the opponent cannot impose excessively large bets (although the explicit inclusion of this kind of regulation in the definition would lead to a hybrid and tediously wordy exposition).

3.6.3. A defect of the first criterion is, in any case, the intervention of an 'opponent': this can make it difficult to avoid the risk, or at least the suspicion, of other factors intruding (such as the possibility of taking advantage of differences in information, competence, or shrewdness). By and large, such possibilities are in 'his' favour (he being the one who decides how much to bet, and in which direction; especially if he is the same person who has chosen the events for which the evaluation of probability is required). Sometimes, however, they can be in your favour (for instance, if, imagining the opponent to have a very distorted opinion, You enunciate an evaluation which induces him to stipulate a bet in a way that You judge favourable[20]).

3.6.4. Under the second criterion these negative features are not present (apart from that inherent in thinking of the various bets as summable). However, given that by choice of the coefficients k_i one can arrange the sizes of the penalties in whatever way one considers most appropriate, even this consequence of 'rigidity' becomes practically negligible. Observe, on the other hand, that the (arbitrary) choice of such coefficients – that is of the *metric* – has no influence at all on the implications of the criterion, since these are always based on merely *affine* notions: the notion of barycentre, and therefore its property of yielding a minimum for the moment, remains invariant under whatever metric one introduces for other purposes, and which occurs in the definition of the moment.

Another doubt arises: one might ask whether there is any good reason for considering the minimization of a penalty L, rather than the maximization of a prize $K - L$. Formally, there is no difference, but if one wants to fix K greater than any possible value of L (in order that the 'prize' always turns out to be positive) one is faced with an annoying limitation (which is impossible anyway if X is not bounded). There is, moreover, an historical reason: when introducing a similar theory in statistical applications, Wald found it natural to posit a Loss in the case of 'wrong decisions' (zero for correct decisions). Finally, one might exercise more care in attempting to prudently minimize a loss (which, in any case, involves uneasiness and disappointment), than in assuring oneself, in a reasonable manner, of the highest level of gain (in this context, the temptation to take a chance is often irresistible; naturally enough, since one cannot lose whatever happens).

3.6.5. One further remark, which is so deeply rooted in what we have said over and over again about the subjective meaning of probability that it is perhaps unnecessary. The two criteria are operational in the sense that they provide a means by which the opinions that an individual carries within himself, whatever they may be, turn out to be observable from the outside. There is no connection with questions like 'what is the *true* value of the probability?' – a question whose meaning finds no place within the present

20 On the eve of a certain football match, You attribute a probability of 40% to the victory of team A, but You think that your opponent, being a supporter of team A, evaluates it at 70%. You can then enunciate a probability of 65%, confident that he will hasten to pay 65 for that which to You is worth 40, but to him 70. But be careful! If he evaluates the probability at 50% instead, and decides to bet in the opposite way, he will pay You 35 for that which is worth 60 to You, and 50 to him – not 30 as You thought.

formulation, and whose meaning I was unable to discover in the attempts made by other theories to provide one – and not even with questions like 'how well founded, or how reasonable, are certain evaluations and their associated motivations?'.

This last question can, in a certain sense, be dealt with by reflecting on various problems that will present themselves from time to time as we proceed to study probability, and to examine various attitudes to both concrete applications and conceptual questions. However, it is mainly a question of arguments (of a rather psychological nature) concerning the choice of a single prevision, **P**, from among the infinite set of coherent previsions, $\mathscr{P}$ (which are equally acceptable from the mathematical point of view). The question does not concern mathematics, except in that it may give a still more enriched description of the various aspects of each choice, so that such a choice, always made absolutely freely, can be made by each individual after an accurate and straightforward examination of everything that in his personal judgment appears relevant for his decision.

3.7 Prevision in the Case of Linear and Nonlinear Dependence

3.7.1. Let us return to the two examples already introduced (Chapter 2, Section 10) under the guise of the logic of certainty. They lend themselves not only to illustrating in practice the application of the two criteria and the consequences of the properties we have established, but above all to developing necessary and instructive insights of a general character. First of all, one notes the essential connection between *linearity* and *prevision* (and the way in which this makes inapplicable to prevision certain arguments which would be valid for prediction). In this connection, it will become clearer how and why it is appropriate to extend the linear ambit $\mathscr{A}$ in relation to the questions to be examined (see the brief explanations given in Chapter 2, Section 2.8).

In the case of a ballot with n voters, we denoted by X, Y, Z, the number of votes cast in favour, against or abstaining, and considered in addition the difference and the ratio of votes for and against, which we denote by $U = X - Y$, $V = X/Y$.

If invited to make a prevision of the outcomes – on the basis of the first or second of the criteria put forward – you provide values $\overline{x},\overline{y},\overline{z},\overline{u},\overline{v}$. These are based on your knowledge, information, impressions or conjectures, about the inclinations or moods of the voters. If your values constituted a prediction, or if You intended to put them forward as *sure*, they would have to be chosen as *integers*, satisfying $\overline{x}+\overline{y}+\overline{z}=n$, $\overline{u}=\overline{x}-\overline{y}$, $\overline{v}=\overline{x}/\overline{y}$. In a prevision they might not even be integers: would the three relations hold? Is it valid to argue that they necessarily hold because they must hold for the true values? In fact, it might seem completely obvious that if the votes for and against *are* x and y (in reality, or in a prediction, or an estimation, or any prevision whatever) then their difference *is* $x - y$, and their ratio *is* x/y, whereas the number abstaining *is* $n - x - y$. For the two linear relations, given the linearity of **P**, this is certainly true, and, considering the ambit $\mathscr{A}$ of (x, y, u) – respectively, of (x, y, z) – is easily interpreted as follows: $\mathscr{Q}$ is the set of points with non-negative integer coordinates x and y with sum $\leqslant n$ in the plane $u = x - y$ – respectively, in the plane $z = n - x - y$ – and a barycentre of such points (with arbitrary weights) can only be some point in the given

plane, in the triangle defined by $x \geqslant 0$, $y \geqslant 0$, $x + y \leqslant n$. If, however, we consider the $\mathscr{A}$ of (x, y, v), the points $\mathscr{Q}$ are those having the same x, y, but now on the surface (hyperbolic paraboloid) $v = x/y$, and the conclusion is no longer valid.[21]

In the case of a random prism with sides X, Y, Z, we denoted the diagonal by U, the area by V, and the volume by W. Since these are not linear functions of the sides, one would not expect that, having evaluated the previsions of the three sides as $\overline{x}$, $\overline{y}$, $\overline{z}$, those of the other elements, say $\overline{u}$, $\overline{v}$, $\overline{w}$, would satisfy the same relations as those holding between the true magnitudes. In other words, in the (six-dimensional) linear ambit $\mathscr{A}$ of (x, y, z, u, v, w), in which $\mathscr{Q}$ is the three-dimensional manifold with equations $u^2 = x^2 + y^2 z^2$, $v = 2(xy + yz + zx)$, $w = xyz$ (with x, y, z, u positive), one would not necessarily expect a barycentre of points of $\mathscr{Q}$ to lie in this manifold. It could be any point $\mathbf{P}$ whatsoever of $\mathscr{P}$, the convex hull of the given $\mathscr{Q}$: once we have evaluated $\overline{x}$, $\overline{y}$, $\overline{z}$, the previsions $\overline{u}$, $\overline{v}$, $\overline{w}$ can only turn out to be some point of the intersection of $\mathscr{P}$ with $x = \overline{x}$, $y = \overline{y}$, $z = \overline{z}$.

If one were interested in a complete solution to the problem, it would be necessary to determine $\mathscr{P}$, or this intersection with it. More generally, one should consider the same problem with certain restrictions on $\mathscr{Q}$: for instance, we might we aware of restrictions like $a' \leqslant X \leqslant a''$, $b' \leqslant Y \leqslant b''$, $c' \leqslant Z \leqslant c''$, or $d' \leqslant X + Y + Z \leqslant d''$, $X \leqslant Y \leqslant 2X$, $Y \leqslant Z \leqslant 2Y$, or that only integer values are admissible for X, Y, Z, or that there are only a finite number of values (x_i, y_i, z_i), and so on. For the purpose of illustration, a simpler version will do: let us suppose that Z is known, $Z = a$ say, and let us consider the restrictions that, given $\overline{x}$ and $\overline{y}$, result for $\overline{u}, \overline{v}$ and $\overline{w}$, *separately*, instead of jointly. In this way, everything is represented each time in a three-dimensional ambit $\mathscr{A}$, which is directly 'visible'.

In the ambit of (X, Y, U) the possible points $\mathscr{Q}$ lie on the circular hyperboloid $u^2 = a^2 + x^2 + y^2$; in fact, if there are no further restrictions, they are all the points on the 'quarter' $x \geqslant 0$, $y \geqslant 0$ of the sheet $u \geqslant 0$; otherwise, they are a subset of these. The barycentre of masses placed on this surface necessarily falls in the convex region that it encompasses (except in the trivial case where the mass is concentrated at a single point, a case where nothing is really random). In a coherent prevision, the diagonal must therefore *necessarily* be estimated longer than it would be if the lengths of the sides coincided exactly with their respective previsions. In the absence of other constraints, given $\overline{x}$ and $\overline{y}$, all the values lying between that minimum and $a + \overline{x} + \overline{y}$ are in fact admissible for $\overline{u}$: one approaches $a + \overline{x} + \overline{y}$ asymptotically by placing two small masses, $\overline{x}/k$ and $\overline{y}/k$, at the points $(k, 0)$ and $(0, k)$, with the rest at the origin, and then letting k become arbitrarily large.

In the ambit of (X, Y, W) the possible points $\mathscr{Q}$ lie instead on the hyperbolic paraboloid $w = axy$ (on the 'quarter' $x \geqslant 0, y \geqslant 0$). In the absence of other restrictions, the convex hull $\mathscr{P}$ is the entire positive orthant, since the barycentre can lie anywhere in this region. In other words, given $\overline{x}$ and $\overline{y}$, $\overline{w}$ can either coincide with $w = a\overline{x}\,\overline{y}$, or can be less (but bounded below by zero), or can be greater (with no constraint). The two limit-cases can be approached by simply placing all the mass at the origin, with the exception, in the first case, of two masses $\overline{x}/k$ and $\overline{y}/k$ at the points $(k, 0)$ and $(0, k)$, respectively, and, in the second case, a single mass $1/k$ at the point $(k\overline{x}, k\overline{y})$. The case $\overline{w} = a\overline{x}\,\overline{y}$ occurs,

21 In this example, provided we do not evaluate the probability that $Y = 0$ as *zero*, we actually have $\overline{v} = \mathbf{P}(V) = +\infty$. This is of more use in showing how one can encounter, in a natural way, cases where the hypothesis of boundedness does not hold, than in illustrating the proposition, which will be clearer after the following example.

for example, under a very important assumption – that of stochastic independence – which we are not yet in a position to discuss. The case of V reduces immediately (having set $Z = a$) to that of W; in fact, $2(XY + aX + aY) = 2W/a + 2a(X + Y)$.

The different behaviour in the two cases we looked at is due to that fact that the points of the first surface were all elliptic, always presenting the concavity in the same direction (delimiting in this way its convex hull), whereas for the second surface, whose points were all hyperbolic, the convex hull of each of its parts is necessarily formed of two parts, adhering to the two faces.

3.7.2. *Functional dependence and linear dependence*. In this context, the representation we have already introduced (Chapter 2, Section 2.8) by means of the dual spaces, $\mathscr{A}$ and $\mathscr{L}$, is appropriate, and provides some insight. $A(X)$ denotes the value which $X = f(X_1, X_2,\ldots, X_n)$ would assume if the X_h (belonging to $\mathscr{L}$) assumed the values x_h (the coordinates of the point A of $\mathscr{A}$): in other words, $A(X) = f(x_1, x_2,\ldots, x_n)$. This is only meaningful if A is one of the possible points $\mathscr{Q}$ of $\mathscr{2}$: that is the values x_h of the X_h are not incompatible. However, we now know that the other points in A – that is the points P of $\mathscr{P}$ the convex hull of $\mathscr{Q}$ – can be interpreted as previsions,[22] and one might ask whether $P(X)$ (understood as above, with $P = A$) is actually the prevision of X. It is clear that this only holds if X belongs to $\mathscr{L}$; in other words, if it is a linear function in the ambit $\mathscr{A}$, or, alternatively, if X is given not just by any function f of the X_h, but in fact by a linear function $X = \sum u_h X_h (h = 0,1,\ldots,n)$. The extension is only valid in this case, and that is why we always confine ourselves to using the notation $A(X)$. The above considerations give us another way of exhibiting the importance and the compass of the linearity. A point P (an admissible prevision) can be either a Q (that is a possible point) in the linear ambit $\mathscr{A}$, or a barycentre of possible points. The knowledge of the barycentre is sufficient, however, to determine only those things which remain invariant under any choice of the points Q and distribution of mass over them so long as one keeps the barycentre fixed.

In other words; one has always to recall that a **P**, defined on any set of random quantities X whatsoever (or, in particular, on any set of events $\mathscr{E}$), is uniquely extendible – and therefore defined – only on the linear space $\mathscr{L}$ of the (finite) linear combinations of $\mathscr{X}$; or, dually, in the corresponding linear ambit $\mathscr{A}$. If $\mathscr{L}$ (or $\mathscr{A}$) is enlarged, one can determine **P** more precisely by more or less arbitrary extensions. So long as we remain in a given ambit $\mathscr{A}$, each point **P** represents, in a manner of speaking, all the **P*** in some larger ambit which have **P** as their projection onto $\mathscr{A}$. This also holds in the infinite-dimensional case, but we postpone any explicit discussion until the Appendix. In order to clear up the simplest cases – one- or two-dimensions – it is sufficient to recall the examples already given in Section 3.7.1 and to examine these further aspects in that context.

3.7.3. *Conclusion*. We conclude, therefore, that whereas it is well known that coherent previsions *preserve* linear dependence, this *only* happens, in fact, *in this case*. In any other case it does not (unless by chance, or under suitable additional hypotheses) because the barycentre of masses lying in a given manifold need not itself belong to the

22 Moreover, it is possible to see that it can even be meaningful to consider an $A(X)$ where A does not belong to $\mathscr{Q}$, and not even to $\mathscr{P}$. For example, one might be interested in the difference between two $\mathscr{P}$, $A(X) = \mathbf{P}_1(X) - \mathbf{P}_2(X)$, and $A = \mathbf{P}_1 - \mathbf{P}_2$ certainly does not belong to $\mathscr{P}$ since we have $A(1) = 0$ instead of $A(1) = 1$.

manifold (except in the trivial case of linearity). In fact, it is sometimes impossible for it to do so (if $\mathcal{Q}$ is the boundary of a convex region).[23]

It is important to always bear in mind details of this kind and to reflect upon them. This is not only, and not mainly, because of their intrinsic importance – however notable this may be – but above all because one has to learn to free oneself from the ever present danger of confusing *prevision* and *prediction*. In a prediction any dependence should obviously be preserved (because it reduces to the choice of *a point* in $\mathcal{A}$, and not the barycentre of masses distributed over $\mathcal{Q}$). The type of argument which, in the examples given, turned out to be wrong if applied to prevision, would, on the other hand, be valid for a prediction. Despite knowing, and remembering, that the arguments do not hold for prevision, anyone (even You, even I) can inadvertently fall into error, applying them without sufficient thought in some particular problem, or in some small corner of the formulation of some particular problem.

There will be many and frequent occasions to warn against errors, misunderstandings, distortions, obscurities, contradictions and the other endless troubles that are so difficult to avoid when dealing with probability, and which are always essentially the result of ignoring the same warning: *prevision is not prediction*! It would not be a bad idea to imagine constantly in front of you an admonitory card – as is used by a certain well-known organization – bearing the message, 'Think!', but with an explanatory rider suited to the needs of probability theory and its applications:

'*Think* : *prevision is not prediction*!'

There is an anecdote, concerning another such maxim, which may perhaps serve to make this recommendation more forceful. It reveals the fallacy of resorting to the self-deception of 'accepting for certain' the alternative on the basis of which one 'decides to act'; a vain attempt to replace a meaningful probability argument by an impossible translation of it into the inadequate logic of certainty. The anecdote is related by Grayson (on p. 52 of a book concerning which we shall have more to say: Chapter 5, 5.5.3) in the following way:

Holes that are going to be dry shouldn't be drilled

'is printed on a sign hanging in one operator's office. This would truly be a "golden rule" if any oil or gas firm could live by it. Unfortunately, no one can – not even this particular operator who drilled 30 consecutive dry holes a few years ago.'

3.8 Probabilities of Events

3.8.1. The properties of probabilities of events are simply special cases of the properties of previsions of random quantities. It will be sufficient to establish them quickly, and to illustrate their meaning within the form of representation that we have introduced.

23 If we wished to be precise, we should exclude the points on the boundary where one does not have strict convexity: in other words, those which are barycentres of other points; or, alternatively, those through which there is no hyperplane which leaves *all* the other points of $\mathcal{Q}$ on the same side.

The theorem of 'total probability'. This is the name given to the theorem that translates, into the field of probability, the additive property of prevision.

The case of incompatibility. If two events A and B are incompatible then, as we have already noted, their logical and arithmetic sums coincide: $E = A \vee B = A + B$, so that, if $\mathbf{P}$ $\mathbf{P}(E) = \mathbf{P}(A) + \mathbf{P}(B)$. The same result holds for the (logical and arithmetic) sum of any finite number of incompatible events: $E = E_1 \vee E_2 \vee \ldots \vee E_n = E_1 + E_2 + \ldots + E_n$, and hence

$$\mathbf{P}(E) = \mathbf{P}(E_1) + \mathbf{P}(E_2) + \ldots + \mathbf{P}(E_n).$$

We can state this formally:

Theorem. *In the case of incompatible events, the probability of the event-sum must be equal to the sum of the probabilities.*

The case of (*finite*) *partitions*. In particular, for a partition in which, in addition, the sum $E = 1$, and hence $\mathbf{P}(E) = 1$, one has the following:

Theorem. *In a* (*finite*) *partition the probabilities must sum to* 1.

In particular, for two complementary events E and $\tilde{E}$ (a partition with $n = 2$), it turns out that $\mathbf{P}(E) + \mathbf{P}(\tilde{E}) = 1$; that is to say, $\mathbf{P}(\tilde{E}) = 1 - \mathbf{P}(E) = \sim\mathbf{P}(E)$; or, in yet another form, if $\mathbf{P}(E) = p$, then $\mathbf{P}(\tilde{E}) = \tilde{p}$.

In words:

Theorem. *The probabilities of two complementary events must themselves be complementary.*

Recalling the properties of the constituents, one can state immediately the following:

Corollary. *In order that the probabilities of all the events E which are linearly dependent on $E_1, \ldots, E_n$ should be determined, it is necessary and sufficient to attribute probabilities to all the constituents $C_1 \ldots C_s$. These probabilities must sum to* 1; *the* $\mathbf{P}(E)$ *depend linearly upon them.*

3.8.2. *Sufficiency of the conditions*. The preceding statements tell us how 'we must' – or 'You must' – evaluate probabilities; in other words, they impose necessary conditions for coherence. In fact – with the obvious restriction that the probabilities be non-negative – they are also sufficient, in the sense that an evaluation satisfying them is coherent, no matter how You choose it. We have already seen this in general in Section 3.4; it may be useful to repeat the argument in this particular case where it is very simple and clear.

Suppose that to the events $E_1 \ldots E_n$ of a finite partition You have attributed non-negative probabilities $p_1 \ldots p_n$, summing to 1, and that I (thinking in terms of 'The first criterion' of Section 3.3) try to force You into a bet which assures me of certain gain. I have to fix the amounts c_i for the bets on the individual E_i in such a way that the resulting bet

$$X = c_1(E_1 - p_1) + c_2(E_2 - p_2) + \ldots + c_n(E_n - p_n)$$

is certainly positive; in other words,

$$c_1E_1 + c_2E_2 + \ldots + c_nE_n > c_1p_1 + c_2p_2 + \ldots + c_np_n$$

no matter which of the E_i occurs. If E_i occurs, however, the left-hand side has the value c_i, and it is impossible for this to be always greater than the right-hand side, since the latter is itself a weighted average of the c_i.

3.8.3. *The case of compatibility; inequality.* For any arbitrary set of events, that is without making the assumption of incompatibility, we have

$$E = E_1 \vee E_2 \vee \ldots \vee E_n = 1 \wedge \left(E_1 + E_2 + \ldots + E_n\right) \leqslant E_1 + E_2 \ldots + E_n$$

and hence

$$\mathbf{P}\left(E\right) \leqslant \mathbf{P}\left(E_1\right) + \mathbf{P}\left(E_2\right) + \ldots + \mathbf{P}\left(E_n\right). \tag{3.6}$$

Stated formally:

Theorem. *The probability of the event-sum must be less than or equal to the sum of the probabilities.*

This is even more evident if one puts it in the form

$$\mathbf{P}\left(E\right) \leqslant \mathbf{P}\left(E_1 + E_2 + \ldots + E_n\right);$$

that is that the probability of the event-sum must be less than or equal to the prevision of the number of successes (one only has to consider that the latter takes into account multiplicities, whereas the former does not).

Expressions in terms of products. In the case of compatible events nothing can be said about $\mathbf{P}(E)$ other than the preceding inequality based just on the $\mathbf{P}(E_i)$. If we introduce other elements, and evaluate them, then, of course, things change. In terms of constituents, the only one we require is

$$C = \tilde{E}_1 \tilde{E}_2 \ldots \tilde{E}_n \quad \text{because} \quad E = \tilde{C}, \quad \mathbf{P}\left(E\right) = 1 - \mathbf{P}\left(C\right).$$

Making use of the products of the E_i (two at a time, three at a time, etc.), and the expansion

$$E = \sum_i E_i - \sum_{ij} E_i E_j + \sum_{ijh} E_i E_j E_h - \ldots \pm E_1 E_2 \ldots E_n, \tag{3.7}$$

we have at once the following:

Theorem. *For the probability of the event-sum we must always have*

$$\mathbf{P}\left(E\right) = \sum_i \mathbf{P}\left(E_i\right) - \sum_{ij} \mathbf{P}\left(E_i E_j\right) + \sum_{ijh} \mathbf{P}\left(E_i E_j E_h\right) - \ldots \pm \mathbf{P}\left(E_1 E_2 \ldots E_n\right). \tag{3.8}$$

Observe that the expression is linear in the probabilities of the products. Note also the special cases of two and three events:

$$\mathbf{P}\left(A \vee B\right) = \mathbf{P}\left(A\right) + \mathbf{P}\left(B\right) - \mathbf{P}\left(AB\right),$$
$$\mathbf{P}\left(A \vee B \vee C\right) = \mathbf{P}\left(A\right) + \mathbf{P}\left(B\right) + \mathbf{P}\left(C\right) - \mathbf{P}\left(AB\right) - \mathbf{P}\left(BC\right) - \mathbf{P}\left(AC\right) + \mathbf{P}\left(ABC\right).$$

3.8.4. *Extensions*. The same formula serves to express the probability that out of n events a given h occur, and no others; and hence the probability that exactly h occur (no matter which ones). The occurrence of $E_1E_2 \ldots E_h$, and no others, can be written as:

$$\begin{aligned} &E_1E_2\ldots E_h\left(1-E_{h+1}\right)\left(1-E_{h+2}\right)\ldots\left(1-E_n\right)\\ &=E_1E_2\ldots E_h-\sum_i E_1E_2\ldots E_hE_{h+i}+\sum_{ij}E_1E_2\ldots E_hE_{h+i}E_{h+j}-\ldots\pm E_1E_2\ldots E_n \end{aligned} \tag{3.9}$$

(where, as can be seen, the sum with k indices is the sum of the products of $h + k$ events; the given h together with k of the others). The event $Y = h$, the number of successes $= h$, is the sum of $\binom{n}{h}$ events of the above kind; in other words, the sum of all the corresponding expressions. In this sum, the products h at a time appear only once, those $h + 1$ at a time appear $h + 1$ times (once for each combination h at a time of their $h + 1$ factors), and so on; in general, the products $h + k$ at a time each appear $\binom{h+k}{h}$ times. For this reason, denoting the sum of the products r at a time by $\Sigma(r)$ for convenience, we have

$$\begin{aligned} \left(Y=h\right)&=\sum{}^{(h)}-\binom{h+1}{h}\sum{}^{(h+1)}+\binom{h+2}{h}\sum{}^{(h+2)}-\ldots\pm\binom{n}{h}\sum{}^{(n)}\\ &=\sum_{r=h}^{n}(-1)^{r-h}\binom{r}{h}\sum{}^{(r)}. \end{aligned} \tag{3.10}$$

If in place of the $\Sigma(r)$ we substitute the sum *of the probabilities* of the products, $\mathbf{P}(E_{i_1}E_{i_2}\ldots E_{i_r}) = p_{i_1i_2\ldots i_r}$, which we denote by S_r for short, the same formula gives the probability

$$\mathbf{P}\left(Y=h\right)=\sum_{r=h}^{n}(-1)^{r-h}\binom{r}{h}\sum p_{i_1i_2\ldots i_r}=\sum_{r=h}^{n}(-1)^{r-h}\binom{r}{h}S_r. \tag{3.11}$$

Note in particular:

$$\begin{aligned} &\mathbf{P}\left(Y=0\right)=1-S_1+S_2-S_3+\ldots\mp S_{n-1}\pm S_n\\ &\mathbf{P}\left(Y=1\right)=S_1-2S_2+3S_3-4S_4+\ldots\mp\left(n-1\right)S_{n-1}\pm nS_n\\ &\mathbf{P}\left(Y=2\right)=S_2-3S_3+6S_4-10S_5+\ldots\mp\binom{n-1}{2}S_{n-1}\pm\binom{n}{2}S_n\\ &\qquad\qquad\qquad\qquad\ldots\\ &\mathbf{P}\left(Y=n-1\right)=S_{n-1}-nS_n\\ &\mathbf{P}\left(Y=n\right)=S_n \end{aligned}$$

(where $\pm$ stands for $(-1)^n$, and $\mp$ for $-(-1)^n$).

Example. A classical and instructive problem is that *of matching*, which lends itself to amusing formulations. If one has n letters and their respective envelopes, what is the probability that if *the letters are inserted into the envelopes at random* one has none, or one, or two, …, or n '*matchings*'; that is letters in their own envelopes? The same problem arises if one pairs up at random right and left shoes from n pairs, or the husbands and wives of n couples, or the jackets and trousers of n suits, and so on. Alternatively, if one

gives back at random to n people their passports, the keys of their hotel rooms, hats left in the cloakroom, and so on. More standard versions are given by the matchings in position among playing cards from two identical decks (for instance by placing them at random in two rows), or between the number of the drawing from an urn of numbered balls and the number of the ball drawn.

The probability of a matching at any given position is obviously $1/n$, of two matchings at two given positions is $1/[n(n-1)]$, and, in general, of r matchings at r given positions is

$$\frac{1}{\left[n(n-1)\ldots(n-r+1)\right]}=\frac{(n-r)!}{n!}$$

(in fact: only one out of the n objects, or only one out of the $n(n-1)$ pairs, ..., or only one out of the $n!/(n-r)!$ arrangements r at a time, is favourable)

The S_r are therefore the sum of $\binom{n}{r}$ terms all equal to $(n-r)!/n!$, so that $S_r = 1/r!$ (independent of n), from which, denoting the number of matchings by Y, we obtain

$$\mathbf{P}(Y=0)=\left\{1-1+\frac{1}{2!}-\frac{1}{3!}+\ldots\pm\frac{1}{n!}\right\}=\mathrm{e}^{-1}-R_n\cong\mathrm{e}^{-1}$$ [24]

$$\mathbf{P}(Y=h)=\left\{1-1+\frac{1}{2!}-\frac{1}{3!}+\ldots\mp\frac{1}{(n-h)!}\right\}\Big/h!=\left(\mathrm{e}^{-1}-R_{n-h}\right)/h!\cong\mathrm{e}^{-1}/h!$$

(in particular: $\mathbf{P}(Y = n-1) = 0$, $\mathbf{P}(Y = n) = 1/n!$). Expressed numerically, $\mathrm{e}^{-1} = 0{\cdot}367879$.

In the limit, as n increases, the distribution tends to that in which

$$\mathbf{P}(Y=h)=\mathrm{e}^{-1}/h!$$

(as we shall see later, Chapter 6, 6.11.2, this is the Poisson distribution with prevision $\mathbf{P}(Y) = 1$).

Observe that for the matching problem one could establish immediately by a direct argument that $\mathbf{P}(Y) = 1$ (i.e. that, in prevision, there is only one matching, whatever n is). We have only to note that it is given by $\mathbf{P}(Y) = n.(1/n)$, since the prevision (probability) of a matching at any one of the n places is $1/n$.

Observe also that the relation $\mathbf{P}(Y = n-1) = 0$ is obvious: in fact, $n-1$ is not a possible value for Y because if we have matchings in $n-1$ positions the last one cannot fail to give a matching (it is as well to point out this fact since it is easily overlooked!).

3.8.5. *Entropy*. Given a partition into events with probabilities $p_1, p_2, \ldots, p_n$, we define the entropy to be the number

$$\sum_h p_h\left|\log_2 p_h\right| \quad \left(\log_2 p_h=(\log p_h)/(\log 2)\right),$$

24 R_n is the remainder of the series $\Sigma\pm1/k!$ from the term $\pm1/(n+1)!$ onwards: it is approximately equal to this first omitted term (which exceeds it in absolute value). With respect to e^{-1} it is practically negligible, except when n (respectively $n-h$) is very small (even for $n = 10$ or $n-h = 10$, the correction does not affect the decimal expression given for e^{-1}).

which represents the prevision of the number of YES–NO questions required to identify the true event.

This is immediate in the case of $n = 2^m$ equally probable events: m YES–NO questions are necessary and sufficient to know with certainty to which half, quarter, eighth,..., of the partition the true event belongs, and finally to know precisely which one it is. If we have nine events with probabilities $\frac{1}{12}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{64}, \frac{1}{64}$, one question suffices if the first one is true; if not (with probability $\frac{1}{2}$) another two questions are sufficient to decide which one of the next three is true, or whether the true event is one of the remaining five; finally (with probability $\frac{1}{8}$), another two questions are necessary, plus (with probability $\frac{1}{32}$) a further one to decide between the last two events. The entropy in this example is therefore given by

$$1 + 2\left(\tfrac{1}{2}\right) + 2\left(\tfrac{1}{8}\right) + \left(\tfrac{1}{32}\right) = 2\tfrac{9}{32} = 2 \cdot 28.$$

If it is not possible to proceed by successive halvings, some fraction is wasted (unless some device is available: for the time being, however, this brief introduction will suffice).

The unit of entropy is called a *bit* (contraction of '*binary digit*'): in the example above, the entropy was 2·28 bits; in the case of $1024 = 2^{10}$ equally probable cases it is 10 bits. For a given n, the entropy is maximized by an equipartition ($p_h = 1/n$): the reader might like to verify this as an exercise.

An item of information that leads to the exclusion of certain of the possible outcomes causes a decrease in entropy: this decrease is called the *amount of information*, and, like the entropy, is measured in bits (it is, in fact, the same thing with the opposite sign: some even call it *negative entropy*). We note that, for the time being, we are not in a position to provide a complete explanation of our assertion that an increase in information causes a decrease in entropy.

3.8.6. *Probability as measure or as mass.* In the set-theoretic interpretation of the events, it appears natural to think of probability – a non-negative, additive function taking the value 1 on the whole space – either as a *measure*, or as a *mass*.

The most widely used approach at the present time is the systematic identification of events as sets, and probability as measure (with all the advantages – as well as the risks! – that derive from a mechanical transposition of all the concepts, procedures and results of measure theory into the calculus of probability). To those reservations that we have already repeatedly expressed in connection with the systematic adoption of the set-theoretic interpretation of events, others must be added (in our opinion) concerning the further inflexibility introduced by the identification of probability with measure. This can, in fact, lead one to think that the representation in a space furnished with a measure binds events and random entities inseparably to a well-determined evaluation of probability. In the most elementary case, where we use Venn diagrams, the figures should be drawn in such a way that the area of each section be equal to its probability (taking the basic rectangle to have unit area). This, on the other hand, is in accordance with those points of view in which to each event (set) there corresponds an objectively (or, in any case, uniquely) determined probability.

If, instead, one wishes to distinguish between, on the one hand, the representation of the logical situation and, on the other hand, the introduction of whatever coherent evaluation of probability one wants to make, it turns out to be preferable to think of

probability as *mass*. The mass can, in fact, be distributed at will, without altering the geometric support and the 'measure', which might in that context appear more natural.[25] In the Venn diagram, without changing the figure in any way, there is no difficulty in imagining the possible ways of distributing a unit mass among the various parts (it does not matter if we put large masses on small pieces) and in imagining those ways that various individuals, real or hypothetical, would have chosen as their own opinion, or those we think they might choose.

Another advantage is the following: if one gives to the space of the representation the structure of the linear ambit, $\mathscr{A}$, then the well-known implications of the mechanical meaning of *mass* make clear all those probabilistic properties that can be translated in terms of knowledge of the barycentre of a distribution (as we have already had occasion to see), or of moments of inertia, and so on.[26]

We shall see shortly some particularly significant applications of this concept in the linear ambit determined by n events (in the sense explained in Chapter 2, 2.8, where the 'possible points' are finite in number, since they correspond to the constituents). Meanwhile, before concluding these comments on the set-theoretic interpretation, it is perhaps instructive to point out the simple, but not obvious, meaning that the expression of the probability of the event-sum acquires under this interpretation. We will consider the case of three events, where

$$E = A \vee B \vee C = A + B + C - AB - AC - BC + ABC.$$

In the Venn diagram, Chapter 2, Figure 2.1b, the area of the union of the pieces A, B, C (or, alternatively, the mass contained in them) is calculated in the following way: firstly, we sum the areas of A, B and C; in this way, however, those of $AB\tilde{C}, A\tilde{B}C, \tilde{A}BC$ (doubly shaded) are counted twice, and that of ABC (triply shaded) is counted three times; subtracting those of AB, AC and BC, we re-establish the correct contribution for those originally counted twice; however, ABC is now counted three times less (since it belongs to AB and AC and BC) and therefore turns out to be ignored altogether; if we add it in, everything turns out as it should be.

3.9 Linear Dependence in General

3.9.1. The straightforward theorems concerning 'total probability', which we established at the beginning of the previous section, certainly require no further explanations. It is, however, convenient to introduce the use of the representation with the spaces $\mathscr{A}$ and $\mathscr{L}$ by means of the simple cases, before proceeding to others of a less trivial nature.

---

25 It is not that different distributions of 'mass' could not equally well be called different 'measures'. It is, however, a fact that when talking in terms of measure one tends to make of it something fixed, with a special status, whereas when talking in terms of mass there is the physical feeling of being able to move it in whatever way one likes.

26 The suggestion has even been put forward that one could always think just in terms of the mass (or measures, or area) rather than in terms of the original meaning of probability: in this way we avoid the questions and doubts of a conceptual nature to which such a notion of probability can give rise. In general, however, in addition to removing the doubts this would also remove the raison d'être of the problems themselves (unless these only involve formal aspects, capable of being isolated from the context which provides them with meaning and content).

We shall restrict ourselves, in general, to the three-dimensional case, which is the most obviously intuitive: the extension to n dimensions (which we shall occasionally mention) presents no difficulty for the reader who is familiar with such things, whereas for those who lack this familiarity it is better to be clear about the simpler case than to acquire confused and formal notions in a less accessible field.

Let E_1, E_2, E_3 be three events (which, for the moment, we take to be logically independent; we shall introduce various assumptions as we go on), and let (x, y, z) be the Cartesian reference system on which we superpose the linear ambit $\mathscr{A}$ and the linear space $\mathscr{L}$. The eight vertices of the unit cube

$$(0,0,0)(1,0,0)(0,1,0)(0,0,1)(0,1,1)(1,0,1)(1,1,0)(1,1,1),$$

thought of as points of $\mathscr{A}$, represent the constituents Q_i forming $\mathscr{Q}$;

$$\begin{array}{cccccccc} Q_0 = & Q_1 = & Q_2 = & Q_3 = & Q_1' = & Q_2' = & Q_3' = & Q_0' = \\ \tilde{E}_1\tilde{E}_2\tilde{E}_3 & E_1\tilde{E}_2\tilde{E}_3 & \tilde{E}_1E_2\tilde{E}_3 & \tilde{E}_1\tilde{E}_2E_3 & \tilde{E}_1E_2E_3 & E_1\tilde{E}_2E_3 & E_1E_2\tilde{E}_3 & E_1E_2E_3 \end{array}$$

(where negations correspond to the *zeros*, affirmations to the *ones*); thought of as points (or vectors) of $\mathscr{L}$, they represent the random quantities

$$0 \quad E_1 \quad E_2 \quad E_3 \quad E_2 + E_3 \quad E_1 + E_3 \quad E_1 + E_2 \quad E_1 + E_2 + E_3$$

(where the presence of a summand corresponds to the *ones*).

The generic point (x, y, z), thought of as a point of $\mathscr{A}$, would mean that E_1 takes the value x, and similarly $E_2 = y$ and $E_3 = z$ (which is invalid, since the random quantities E_i cannot take on values other than 0, 1). This can be valid, however, as *prevision*, in the sense that $\mathbf{P}(E_1) = x$, $\mathbf{P}(E_2) = y$, $\mathbf{P}(E_3) = z$; in other words, (x, y, z) represents the prevision $\mathbf{P}$ which attributes to E_1, E_2, E_3 the probabilities $(p_1, p_2, p_3) = (x, y, z)$, and which is also expressible as the barycentre of the points Q_i with suitable weights (masses) q_i. Thought of as a point (or vector) of $\mathscr{L}$, (x, y, z) represents the random quantity $X = uE_1 + vE_2 + wE_3$ with coefficients $(u, v, w) = (x, y, z)$. Since $\mathbf{P}(X) = up_1 + vp_2 + wp_3 = ux + vy + wz$, $\mathbf{P}(X)$ can be interpreted as the inner product of the (dual) vectors $\mathbf{P}$ (or $P - 0$) of $\mathscr{A}$ and X (or $X - 0$) of L; or, alternatively, as $\mathbf{P}(X) = (P - 0) \times (X - 0)$ in the metric space on which $\mathscr{A}$ and $\mathscr{A}$ have been superposed.

Until we state precisely the assumptions made concerning the E_i, that is establish which among the eight products are actually possible constituents, all this remains rather general and introductory in character; simply a repetition of things we know already, with a few additional details.

3.9.2. *The case of partitions.* If the E_i constitute a partition, there are three constituents. $Q_1 = (1, 0, 0)$, $Q_2 = (0, 1, 0)$, $Q_3 = (0, 0, 1)$. We know that the p_i can be any three non-negative numbers summing to 1. In other words, the admissible $\mathbf{P} = (x, y, z)$ belong to the plane $x + y + z = 1$. More precisely, they belong to the triangle having as its vertices the three possible points Q_1, Q_2, Q_3, and are, in fact, uniquely expressible as barycentres of these points, $P = q_1Q_1 + q_2Q_2 + q_3Q_3$, with weights $q_1 = x$, $q_2 = y$, $q_3 = z$. This triangle constitutes the space $\mathscr{P}$ of admissible previsions, and is precisely the convex hull of the set $\mathscr{Q}$ of possible outcomes (which reduces in this case to the three given vertices). Representing the triangle by a figure in the plane, one sees that the

probabilities x, y, z, turn out to be the barycentric coordinates of the point P with respect to the Q_i. Since the triangle is equilateral, one has the standard 'ternary diagram' (as is used, for example, to indicate the composition of ternary alloys) in which x, y, z also have a more immediate interpretation as the *distances of the point from the sides*, taking as unity the height of the triangle (to which the sum of the three distances is always equal). It is also clear that a point outside of the triangle (not in the plane, or in the plane but outside the triangle) can be brought nearer to all the three vertices – that is to all the points of $\mathscr{Q}$ – by transporting it into the triangle. This can be accomplished by projecting it onto the plane, and then, if the projection falls outside the triangle, by transporting it to the nearest point on the boundary. This is related to the 'second criterion', if we think of the penalty as being the square of the ordinary distance in this representation.

If we think in terms of $\mathscr{L}$, we could say, instead, that the point (1, 1, 1) represents the random quantity that is certainly equal to 1, given that $E_1 + E_2 + E_3 = 1$. The fact that for the coordinates of P we must have $x + y + z = 1$ is then interpreted on the basis of the scalar product: $\mathbf{P}(1) = x \cdot 1 + y \cdot 1 + z \cdot 1 = 1$.

3.9.3. *The case of incompatibility.* If the E_i are incompatible (but not exhaustive) there are four constituents: the previous three and $Q_0 = (0, 0, 0)$; that is Q_0, Q_1, Q_2, Q_3. The above considerations still hold, except that we now have the relation $x + y + z \leqslant 1$ (instead of = 1). We still have P expressible uniquely as a barycentre, $P = q_0Q_0 + q_1Q_1 + q_2Q_2 + q_3Q_3$, of the Q_i, with weights $q_0 = 1 - x - y - z$, $q_1 = x$, $q_2 = y$, $q_3 = z$, and the space $\mathscr{P}$ (which was the triangle with vertices Q_1, Q_2, Q_3) is now the tetrahedron having in addition the vertex Q_0.

3.9.4. *The case of a product.* Let E_1 and E_2 be logically independent, and E_3 be their product: $E_3 = E_1E_2$. The constituents are then the following four: $Q_0 = (0, 0, 0)$, $Q_1 = (1, 0, 0)$, $Q_2 = (0, 1, 0)$ and $Q_0' = (1, 1, 1)$. The first three are in the plane $z = 0$, the last three are on $z = x + y - 1$; the other two groups of three are on $z = y$ and $z = x$, respectively. The space $\mathscr{P}$ is, therefore, the tetrahedron $z \geqslant 0$, $z \geqslant x + y - 1$, $z \leqslant x$, $z \leqslant y$, or, in other words, expressed compactly using $\wedge$ and $\vee$,

$$[\max(0, x + y - 1) =] \qquad 0 \vee (x + y - 1) \leqslant z \leqslant x \wedge y \qquad [= \min(x, y)].$$

These are the restrictions under which one can arbitrarily choose the probabilities of two logically independent events and that of their product.

Here also, P is uniquely expressible as a barycentre

$$P = q_0Q_0 + q_1Q_1 + q_2Q_2 + q_0'Q_0'$$

of the Q with weights $q_0 = 1 - x - y + z, \quad q_1 = x - z, \quad q_2 = y - z, q_0' = z.$

3.9.5. *The case of the event-sum.* This proceeds as above, except that $E_3 = E_1 \vee E_2$ (instead of E_1E_2). Since the event-sum is $E_1 + E_2 - E_1E_2$, this case reduces straightaway to the preceding ones. The constituents are Q_0, Q_1', Q_2', Q_0'; the inequalities for the tetrahedron $\mathscr{P}$ having these vertices are

$$\left[\max\left(x, y\right) =\right] \qquad x \vee y \leqslant z \leqslant 1 \wedge \left(x + y\right) \qquad \left[= \min\left(1, x + y\right)\right];$$

the weights which give $P = (x, y, z)$ as a barycentre in terms of the Q are

$$q_0 = 1 - z, \qquad q_1' = z - y, \qquad q_2' = z - x, \qquad q_0' = x + y + z.$$

Remark. In the preceding cases each P was derived as a barycentre of the Q *with uniquely determined weights q;* it is important to note (and we shall return to this later) that this circumstance is exceptional. To be more precise, this happens when and only when the Q are *linearly independent* – in the examples above we had, in fact, either three noncollinear, or 4 noncoplanar – or when they are (as events) *expressible as a linear combination of the given events*. In fact, they were, in the first case, E_1, E_2, E_3; in the second, $1 - E_1 - E_2 - E_3, E_1, E_2, E_3$; in the third, $1 - E_1 - E_2 + E_3, E_1 - E_3, E_2 - E_3, E_3$; and in the fourth, $1 - E_3, E_3 - E_2, E_3 - E_1, E_1 + E_2 - E_3$. In other words, the Q (as events) belonged in these cases to $\mathscr{L}$. Observe that the expressions for the Q in terms of the E are the same as those for the weights q in terms of x, y, z. *In the following examples this will no longer happen*.

3.9.6. *The case of exhaustivity*. If we specify *only* that E_1, E_2, E_3 are exhaustive, then there are seven constituents; the eight minus $Q_0 = (0, 0, 0)$, which is excluded. This latter vertex of the cube being missing, the convex hull $\mathscr{P}$ is the cube itself minus the tetrahedron defined by this vertex and the three adjacent ones; that is the part of the cube $0 \leqslant x, y, z \leqslant 1$ which satisfies the inequality $x + y + z \geqslant 1$. Each of its points P can be expressed – *in an infinite number of ways* – as a barycentre of points Q (unless the point coincides with a vertex, or belongs to an edge, or a triangular face, in which case the number of representations is finite). In fact, all we have to do is to choose non-negative weights q, summing to 1, such that

$$q_1 + q_2' + q_3' + q_0' = x, \qquad q_2 + q_1' + q_3' + q_0' = y,$$
$$q_3 + q_1' + q_2' + q_0' = z$$

(4 equations and 7 unknowns).

3.9.7. *The case where the negations are also exhaustive*. If we exclude both the extreme constituents, that is in addition to $Q_0 = (0, 0, 0)$ we also exclude $Q_1' = (1, 1, 1)$, then six constituents remain. The cube has now had removed from it the two opposite tetrahedrons, and the remaining part $\mathscr{P}$ is that defined by the double inequality $1 \leqslant x + y + z \leqslant 2$.[27] Other considerations are as above.

A useful example is given by the comparisons between three random quantities, X, Y, Z; in other words, by considering the three events $E_1 = (X > Y)$, $E_2 = (Y > Z)$, $E_3 = (Z > X)$ (we assume excluded, or at least as practically negligible, the case of equality). By transitivity, the three events cannot turn out to be either all true or all false; there remain the other six constituents, corresponding to the $6 = 3!$ possible permutations. As an application, one might think, for example, of comparing the weights (or temperatures, etc.) of three objects.

Other cases. The following cases are similar (and are useful as exercises): $E_3 \subset E_1E_2$ (5 constituents); E_1 and E_2 incompatible, $E_3 \subset (E_1 = E_2)$ (6 constituents); and so on. Another example with four (independent!) constituents is given by $E_3 \equiv (E_1 = E_2)$.

3.9.8. *The case of logical independence*. All eight constituents exist; $\mathscr{P}$ is the whole cube. This is the most complete and 'normal' case; there is little to say apart from thinking about it in the light of remarks concerning more elaborate cases.

---

27 The case $x + y + z = 2$ (with constituents Q_1', Q_2', Q_3') is similar to that of the partition $(x + y + z = 1)$, and is obtained if $\tilde{E}_1, \tilde{E}_2, \tilde{E}_3$ form a partition.

The same case in any number of dimensions. If $E_1, E_2,\ldots, E_n$ are logically independent events, we will have 2^n constituents Q_i, the vertices of the unit hypercube; that is the points $(x_1, x_2,\ldots, x_n)$ in the linear ambit $\mathscr{A}$ with $x_i = 1$ or 0. The admissible previsions **P** are those of the cube $\mathscr{P}$, $0 \leqslant x_i \leqslant 1$, which is the convex hull of the set of the vertices $\mathscr{Q}$. The linear space L is formed by the random quantities $X = u_1E_1 + u_2E_2 + \ldots + u_nE_n$, which are linearly dependent (homogeneously, but it is easy to take into account separately an additive constant) on the events E_i. Conceptually, everything that has been stated for $n = 2$ and $n = 3$ also holds for arbitrary n (this saves us repeating everything in a more cumbersome notation and so making the exposition rather heavy going).

3.9.9. *General comments.* Each particular case differs from the final one by virtue of the exclusion of some of the constituents: instead of 2^n there are only $s < 2^n$. These determine a linear space of dimension $d(d \leqslant n, \log_2 s \leqslant d \leqslant s - 1)$; if $d < n$, the n events E_i are linearly dependent. In fact, if all the Q satisfy a linear relation $\sum_i x_i = \text{const.}$ the same holds for the E_i. For instance, in the above examples $x + y + z = 1, x + y + z = 2$ gives $E_1 + E_2 + E_3 = 1$ (or 2), so that we need only consider two events, for example E_1 and E_2, setting $E_3 = 1 - E_1 - E_2$ or $E_3 = 2 - E_1 - E_2$, respectively (this also holds in the general case). If we consider the unnecessary E_i as eliminated (since they are linearly dependent on the others), we can always arrange that $d = n$; in any case, $\mathscr{P}$ is the convex hull (d-dimensional polyhedron) having as vertices the points Q which form $\mathscr{Q}$.

Given some **P** (in $\mathscr{A}$), in other words, *having evaluated the probabilities* $\mathbf{P}(E_i)$ *of the given events*, **P** turns out to be determined for all those random quantities X which are linearly dependent on the E_i, and for no others; that is for those belonging to $\mathscr{L}$. In particular, the probability of an event E is determined if and only if E is one of these X.[28] This statement takes into account all the obvious cases: for example, the probability of $A \vee B$ is not determined by $\mathbf{P}(A)$ and $\mathbf{P}(B)$ (unless we assume incompatibility), but is determined if we include $\mathbf{P}(AB)$, since we have the relation $A + B = AB + A \vee B$. It is useful to see an example of how nontrivial events can be found among the X of $\mathscr{L}$ (i.e. the X that only have two possible values; which we can always represent as 0 and 1). We shall see then that, if E is not linearly dependent on the E_i, one can only say that $p' \leqslant \mathbf{P}(E) \leqslant p''$, where $p' = \sup \mathbf{P}(X)$ for the X of $\mathscr{L}$ which are certainly $\leqslant E$, and $p'' = \inf \mathbf{P}(X)$ for the X of $\mathscr{L}$ which are certainly $\geqslant E$.

3.9.10. *A nonobvious example of linear dependence.* Suppose that A, B, C, D, F, G are the participants in a competition, and that six other individuals each choose from among the participants their three 'favourites' (a prize being offered to all those who have included the winner among their 'favourites'). Suppose also that we know the choices to be: C, D, G for the first individual; B, C, G for the second; A, D, F for the third; B, F, G for the fourth; A, C, D for the fifth; D, F, G for the sixth. Finally, a seventh individual – suppose it is You – has chosen A, B, C. Is your guess, the event E say, linearly

28 This does not exclude the possibility that for certain evaluations (limit-cases in which an inequality reduces to an equality) $\mathbf{P}(E)$ can turn out to be determined for E which are not linearly dependent on the E_i we started with (and perhaps not even logically dependent). For instance, if neither of A and B is logically dependent on $A \vee B$, then knowing $\mathbf{P}(A \vee B)$ is not sufficient to determine $\mathbf{P}(A)$ and $\mathbf{P}(B)$; if, however, $\mathbf{P}(A \vee B) = 0$, it follows necessarily that $\mathbf{P}(A)$ and $\mathbf{P}(B)$ are also zero.

dependent on the events $E_1,\ldots, E_6$, which denote the guesses of the others, or not? This question might be important, for example, in the following situation: there is an expert in whom You have great confidence, so far as judging the competition and the participants is concerned, and whose opinion concerning your probability of winning is of interest to You. However, You do not know this directly (since You do not know what probability of winning he attributes to each participant) but only indirectly (because You happen to know what probabilities he attributes to the guesses of the others turning out to be correct); is this enough?

We have the system of equations:

$$\begin{aligned}
1 &= A + B + C + D + F + G \\
E_1 &= \quad\quad C + D \quad\quad + G \\
E_2 &= \quad B + C \quad\quad\quad + G \\
E_3 &= A \quad\quad\quad + D + F \\
E_4 &= \quad B \quad\quad\quad + F + G \\
E_5 &= A \quad\quad + C + D \\
E_6 &= \quad\quad\quad\quad D + F + G \\
E &= A + B + C
\end{aligned}$$

(the first equation states, as we have seen, that the six cases are the only ones possible, and are incompatible). One could work out the determinant (and, by virtue of its being zero, could verify linear dependence), but instead we note (and leave the reader to verify it by working out the sum) that we have the relation

$$\begin{aligned}
&2E_1 - E_2 + E_3 - 3E_4 - 5E_5 + 5E_6 + 7E \\
&= 3\big(A + B + C + D + F + G\big) = 3,
\end{aligned}$$

from which

$$E = \frac{1}{7}\big(3 - 2E_1 + E_2 - E_3 + 3E_4 + 5E_5 - 5E_6\big).$$

Hence, if I know the $p_i = \mathbf{P}(E_i)$ of the guesses, I can conclude that in the expert's opinion (assumed coherent) $p = \mathbf{P}(E)$ must be

$$p = \frac{1}{7}\big(3 - 2p_1 + p_2 + p_3 + 3p_4 + 5p_5 - 5p_6\big).$$

In a similar way one could, of course, see whether the $p_1 \ldots p_6$ are admissible (compatible with probabilities $\geqslant 0$ for the partition A, B, C, D, F, G, and if, in any case, they determine them, etc.). It may be a useful exercise to develop these questions in the context of this example (as it stands, or modifying it in some way).

3.10 The Fundamental Theorem of Probability

3.10.1. We turn now to proving and illustrating the general conclusion that we stated before, and which, in a more complete and precise form, constitutes the following:

Theorem. *Given the probabilities* $\mathbf{P}(E_i)$ $(i = 1, 2,\ldots, n)$ *of a finite number of events, the probability,* $\mathbf{P}(E)$*, of a further event E, either*

a) *turns out to be determined* (*whatever* **P** *is*) *if E is linearly dependent on the* E_i (*as we already know*)*; or*
b) *can be assigned, coherently, any value in a closed interval* $p' \leqslant \mathbf{P}(E) \leqslant p''$ (*which can often give an illusory restriction, if* $p' = 0$ *and* $p'' = 1$*, or in limit-cases for particular* **P***, give a well-determined result* $p = p' = p''$).

More precisely, p' *is the upper bound,* sup $\mathbf{P}(X)$*, of the evaluations from below of the* $\mathbf{P}(X)$ *given by the random quantities X of* $\mathscr{L}$ (*i.e. linearly dependent on the* E_i) *for which we certainly have* $X \leqslant E$*. If E is not logically dependent on the* E_i*, observe that* $X \leqslant E$ *can be more usefully replaced by* $X \leq E'$ *where* E' *is the largest event logically dependent on the* E_i *contained in E* (*see Chapter 2, 2.7.3*)*. The same can be said for* p'' (*replacing* sup *by* inf*, maximum by minimum,* E' *by* E''*, and changing the direction of the inequalities, etc.*).

Proof. If $Q_1 \ldots Q_s$ denote the constituents, relative to $E_1 \ldots E_n$, and E is logically (but not linearly) dependent on the E_i, then the linear ambit $\mathscr{A}'$ obtained by the adjunction of E (i.e. by adding a new coordinate x to the preceding $x_1 \ldots x_n$) has the same constituents Q_h, but now placed at the vertices of a cube in $n + 1$ dimensions instead of n. Each $Q = (x_1, x_2,\ldots, x_n)$ is either left as it was (with $x = 0$), or moved onto the parallel S_n $(x = 1)$, becoming either $(x_1, x_2,\ldots, x_n, 0)$ or $(x_1, x_2,\ldots, x_n, 1)$, according to whether Q is contained in $\tilde{E}$ or in E. The convex hull $\mathscr{P}'$ in S_{n+1} (in $\mathscr{A}'$) has as its projection onto the preceding S_n ($\mathscr{A}$) the preceding $\mathscr{P}$. For each admissible **P** in the latter (with coordinates $p_i = \mathbf{P}(E_i)$), the admissible extensions in $\mathscr{A}'$ are the points $\mathbf{P}'$ that project onto **P** and belong to $\mathscr{P}'$; that is. belong to the segment $p' \leqslant x \leqslant p''$ which is the intersection of the ray $(p_1, p_2, \ldots, p_n, x)$ with $\mathscr{P}'$. The extreme points $(x = p', x = p')$ are on the boundary of $\mathscr{P}'$, that is on one of the hyperplanes (in n dimensions) that constitute its faces (they could be on more than one – vertices, edges, etc. – but this does not affect the issue). Suppose the hyperplane is given by $\sum u_i x_i + ux = c$; in other words, suppose that the relation $\sum u_i E_i + uE = c$ holds on it, that is that $E = (c - \sum u_i E_i)/u$: then the X in $\mathscr{L}$ defined by the right-hand side has the given property, and yields $p' = \mathbf{P}(X)$. Similarly for p''.

3.10.2. *Applications.* Let us generalize some of the examples considered previously in S_3. Those concerning the number of successes,

$$Y = E_1 + E_2 + E_3,$$

now become the consideration of $Y = E_1 + E_2 + \ldots + E_n$, and we can look at various subcases. Suppose that either Y is known, $Y = y$ $(0 \leqslant y \leqslant n)$ (as in the previous cases where $Y = 1$ and $Y = 2$), or certainly lies between two given extreme values y' and y'' $(0 \leqslant y' \leqslant y'' \leqslant n)$ (as in the previous cases, where $1 \leqslant Y \leqslant 2$). The interpretation of this last example, as given in Section 3.9.7, will now be extended (in different ways) to comparisons between n objects: finally, the case of the event-sum will require all the products.

3.10.3. *Knowledge about frequency.* This first example is noteworthy in that it constitutes the first and most elementary link in the long chain of conclusions which, as we proceed, will clarify and enrich our insight into the relationship that holds between *probability* and *frequency*. This is important both for what the conclusions do say and,

perhaps even more so (in some situations at least), in order to get used to not interpreting them as saying something which they do not say.

The simplest case is that in which the number of successes, $Y = E_1 + E_2 + \ldots + E_n$, is known (for certain);that is the *frequency Y/n* is known (for certain). Let $Y = y$, so that $Y/n = y/n$. The following are possible examples: in an election, out of n candidates we know that y are to be elected; in an examination, y candidates out of n passed (but we are still ignorant of which ones); in a drawing of the lottery, out of $n = 90$ numbers $y = 5$ will be drawn; at $n = 90$ successive drawings of all the balls in Bingo, all the $y = 15$ numbers on your card will come out.

As an extension, we have the case in which we know the limits between which Y must lie; $y' \leqslant Y \leqslant y''$ (and hence that the frequency must be between y'/n and y''/n). In the preceding examples: it may be that the electoral system allows the number elected to vary between y' and y''; that on the basis of partial information about the examinations one knows that at least y' have passed and at least $n - y''$ have not; if we consider 10 drawings of the lottery instead of one (for instance, all the 10 'wheels' on the same day), then of the $n = 90$ numbers the total of different numbers drawn can vary between $y' = 5$ (all the sets of five identical) and $y'' = 50$ (no number repeated).

It is obvious that, as in the case $n = 3$, the sum of the $\mathbf{P}(E_i)$, that is $\mathbf{P}(Y)$, must give in the first case y, and in the second a value $y' \leqslant \mathbf{P}(Y) \leqslant y''$. Put more forcefully; dividing by n, *the probabilities* $\mathbf{P}(E_i)$ *must be such that their arithmetic mean coincides with the known frequency* y/n, *or falls between the extreme values,* y'/n *and* y''/n, *that the frequency can assume* (end-points included). This is all that can be said on the basis of the given information. In general, one might say more: for example, that each number in the lottery has probability $\frac{5}{90}$ of coming up in a given drawing, and not different probabilities with mean $\frac{5}{90}$. This could only be done, however, on the basis of additional knowledge or considerations which must be kept separate.

3.10.4. *The linear ambit of events logically dependent on n given events.* For the purpose in hand, it is obviously sufficient to consider the linear ambit, let us call it $\mathscr{A}^*$, generated by the s constituents Q_h (these form a partition, and so the dimension is actually $s - 1$, given the identity $Q_1 + Q_2 + \ldots + Q_s = 1$). We could also generate it by means of the E_i and their products (two at a time, three at a time, etc.). We saw, in Section 3.8.3, that in this way one can express the event-sum linearly, and we shall now see that it is possible to express all the constituents linearly, and hence all the events which are logically dependent on the E_i. We will suppose that the E_i are logically independent, so that $s = 2^n$; in the other case, the treatment is equally valid, except that the constituents and the products which turn out to be impossible have to be omitted.

Let us illustrate the situation by referring to the case of three logically independent events and their products; for convenience we denote the three events by A, B, C (instead of E_1, E_2, E_3) and their products by $F = AB$, $G = AC$, $H = BC$ and $E = ABC$. We have seven events that are linearly independent because there exists only one linear relation between the $2^3 = 8$ constituents (the sum =1). Some inequalities (implications) hold among them, however; for instance, $A \geqslant AB \geqslant ABC$ so that $A \geqslant F \geqslant E$ (as is obvious if one considers that of the $2^7 = 128$ vertices of the cube in seven dimensions only the eight corresponding to the constituents relative to A, B, C, are possible).

We list the constituents, giving their coordinates in the ambit $\mathscr{A}^*$, and the linear expressions in the dual space $\mathscr{L}^*$:

$$ABCFGHE = (1,1,1,1,1,1,1) = E,$$
$$AB\tilde{C}F\tilde{G}\tilde{H}\tilde{E} = (1,1,0,1,0,0,0) = F - E,$$
$$A\tilde{B}C\tilde{F}G\tilde{H}\tilde{E} = (1,0,1,0,1,0,0) = G - E,$$
$$\tilde{A}BC\tilde{F}\tilde{G}H\tilde{E} = (0,\ 1,\ 1,\ 0,\ \ 0,\ 1,\ \mathbf{0}) = H - E,$$

$$A\tilde{B}\tilde{C}\tilde{F}\tilde{G}\tilde{H}\tilde{E} = (1,\ 0,\ 0,\ 0,\ 0,0,\ 0) = A - F - G + E,$$
$$\tilde{A}B\tilde{C}\tilde{F}\tilde{G}\tilde{H}\tilde{E} = (0,\ 1,\ 0,\ 0,0,0,\ 0) = B - F - H + E,$$
$$\tilde{A}\tilde{B}C\tilde{F}\tilde{G}\tilde{H}\tilde{E} = (0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0) = C - G - H + E,$$
$$\tilde{A}\tilde{B}\tilde{C}\tilde{F}\tilde{G}\tilde{H}\tilde{E} = (0,\ 0,\ 0,\ 0,\ 0,\ 0,0) = 1 - A - B - C + F + G + H - E.$$

These expressions, and the analogous ones for each of the events logically dependent on A, B, C, are obtained as shown in the following example:

$$\tilde{A}B\tilde{C} = (1 - A)B(1 - C) = B - AB - BC + ABC = B - F - H + E.$$

The necessary and sufficient condition for coherence is that the probabilities of the constituents are non-negative (they automatically turn out to sum to 1), and therefore the following inequalities (where, for simplicity, we denote the probability of an event by the corresponding lower case letter) are necessary and sufficient:

$$e \geqslant 0, \qquad f, g, h \geqslant e, \qquad a \geqslant f + g - e, \qquad b \geqslant f + h - e,$$
$$c \geqslant g + h - e, \qquad (a + b + c) - (f + g + h) + e \leqslant 1.$$

3.10.5. *A canonical expression for random quantities*. By analogy, we indicate here how, in the same manner, each random quantity

$$X = c_0 + c_1E_1 + c_2E_2 + \ldots + c_nE_n,$$

linearly expressible in terms of the events E_i, can be put in a meaningful canonical form by reducing it to a linear combination

$$X = x_1C_1 + x_2C_2 + \ldots + x_sC_s$$

of the constituents C_h (the x_h are the possible values of X, assumed in correspondence to the occurrence of the C_h). As an example: if we denote two logically independent events by A and B, and the constituents by $Q_1 = AB$, $Q_2 = A\tilde{B}$, $Q_3 = \tilde{A}B$, $Q_4 = \tilde{A}\tilde{B}$, where $1 = Q_1 + Q_2 + Q_3 + Q_4$, $A = Q_1 + Q_2$, $B = Q_1 + Q_3$, we have, for instance, for $X = 3 - 4A + B$:

$$\begin{aligned} X &= 3(Q_1 + Q_2 + Q_3 + Q_4) - 4(Q_1 + Q_2) + (Q_1 + Q_3) \\ &= 0 \cdot Q_1 + (-1) \cdot Q_2 + 4 \cdot Q_3 + 3 \cdot Q_4. \end{aligned}$$

X assumes the possible values -1, 0, 3, 4, corresponding to Q_2, Q_1, Q_4, Q_3.

3.10.6. *Comment*. The above considerations are intended to familiarize the reader (in the case of events) with the crucially important idea of the relations of linearity and

inequality, and to stress *a fact* and *a criterion* that will be of use in what follows, and more generally.

The *fact* is the *possibility* of expressing all that can legitimately be said by arguing solely in terms of the events (and random quantities) whose prevision is known. That is to say, without leaving the linear ambit determined by the latter, without imagining already present a probability distribution over larger ambits, those in which the extension is possible, albeit in an infinite number of ways.

The *criterion* lies in the *commitment* to systematically exploiting this fact; the commitment considered as the expression of a fundamental methodological need in the theory of probability (at least in the conception which we here maintain). All this is not usually emphasized.

These considerations should go some way to excusing the length of the exposition, which is certainly excessive in comparison with what would be desirable if this topic were well enough known in general to permit us to restrict ourselves to a few brief remarks.

3.10.7. *The case of an infinite number of events* (*or random quantities*). The fundamental theorem of probability (and prevision), given in Section 3.10.1, permits us – even in countably infinite or nondenumerable cases where, of course, the number of choices is infinite – to proceed to attribute to all the events and random quantities that we wish, one after the other, probabilities and previsions coherent with the preceding ones. The arguments presented do not become invalid when we pass to the infinite case, because the conditions of coherence always refer just to finite subsets: see Appendix, Section 15.

This demonstrates the theorem of the *unconditional existence and extendibility of coherent previsions* of events and random quantities in any (open)[29] field. In other words:

> *If within the field in which they are made, the previsions do not already give rise to incoherence, no incoherence arises to prevent the existence of coherent previsions in any field whatever, coinciding with the preceding ones whenever these apply.*

3.11 Zero Probabilities: Critical Questions

3.11.1. In both the criteria put forward in order to define probability there was a point whose clarification we held over to the sequel. It was the same point in both cases; the wherefore of the precaution taken in excluding the possibilities of gains being *all uniformly negative*, but not that of gains being all negative (without the 'uniformity' condition). Another matter, connected with this, is the removal of the reservations regarding the prevision of unbounded random quantities.

We are dealing with critical questions and, if we only wished to consider those aspects relating to applications, they could be omitted, or confined to the Appendix. This, however, is not possible. In Chapter 6, we have to study distributions, and to throw light on the conceptual differences and their wherefores, introduced in accordance with the

29 'Open' is meant in the sense of not being preconstituted, not constrained, not a 'Procrustean bed', not a Borel field, not consisting of events, etc., that have a given meaning or structure, but a field in which we can, at any moment, insert whatever might come to mind.

present viewpoint, it is better to focus right from the beginning on those aspects which will play a fundamental rôle.

The fact is that a logical construction is such in so far as it is a whole in which 'tout se tient' ['everything fits together']; otherwise, it is nothing of the sort. Questions that are seemingly completely otiose and insignificant can have, and do have, interconnections with all the rest and are essential for an understanding of them. To ignore them, or merely to mention them in passing, is dangerous, especially when they impinge on delicate and controversial matters: too many ideas then remain rather vague and give rise to an accumulation of doubts.

For this reason, having reached the end of Chapter 3, we shall now consider the questions of a critical nature that have arisen; we shall do the same at the end of Chapter 4, coming back to these same questions under a new guise; finally, at the end of Chapter 5, we shall arrive at the same kind of considerations, although with respect to topics that are less technical and more general. We shall attempt to confine ourselves to the minimum necessary discussion, expressed as simply as possible. The few additional clarifications or examples will be recognizable 'at a glance' by virtue of the small print.

3.11.2. It would not be accurate to say that all the problems reduce to the presence of zero probabilities but, in order to have a guideline to follow, it is convenient to think in these terms (just as it is not only suggestive but also appropriate to mention them in the section heading).

It seems impossible that there is anything at all to be said about zero probabilities. Instead, we have the following basic questions:

i) Can a possible event have zero probability? If so:
ii) Is it possible to compare the zero probabilities of possible events (to say if they are equal, or what their relation is, etc.)?
iii) Can a union of events with zero probabilities have a positive probability (in particular, can it be the certain event)?
iv) Are there any connections with problems concerning random quantities, and in particular with the problem of prevision for unbounded random quantities?

Question (II) crops up again within the topics of Chapter 4 and will be discussed there; we had to mention it, not only to put it in its natural position as a 'question' but also to give prior warning that any incidental comments that we make here for convenience will be clarified at the appropriate place: we will draw attention to this by writing '(II!)'.

Questions (I) and (III) can be bracketed and discussed together straightaway; afterwards we shall pass on to (IV). However, there was a reason for putting the two questions (I) and (III) separately. Question (III), which evidently requires to be put in the context of infinite partitions, might lead one to think and state that one can only have possible events with zero probability *if they belong to infinite partitions* (!). This is monstrous. If E has probability p (in particular = 0) it *is* an event with probability p (in particular with zero probability) both when considered in itself, or in the dichotomy E and $\tilde{E}$, or in any other partition into few, many or an infinite numbers of events, obtained by partitioning in any way whatsoever. Unfortunately, this propensity to see each event embedded in some scheme, together with others usually studied with it, gives rise to serious confusions both in theoretical matters (as is the case here) and practically (as in the examples in Chapter 5, 5.8.7).

This having been said as an appropriate warning, we can pose question (III) once again by asking *whether in an infinite partition one can attribute zero probability to all the events*. In this form, the question becomes essentially equivalent to that concerning the different types of additivity: *finite*, only for a finite sum; *countable*, for the denumerable case; *perfect*, if the additivity always holds.

There are precisely three answers, corresponding to these three types (with a variation, which is related to (I)):

A = Affirmative, N = Negative (N' and N''), C = Conditional

(and in what follows, we shall denote them and the corresponding points of view with the initials A, N and C, or, if necessary, A, N', N'' and C).

A: Yes. Probability is finitely additive. The union of an infinite number of incompatible events of zero probability can always have positive probability, and can even be the certain event.

N: No. Probability is perfectly additive. In any partition there is a finite, or countable, number of events with positive probabilities, summing to one: the others have zero probability both individually and together.

C: It depends. The answer is NO if we are dealing with a countable partition, because probability is countably additive; the sum of a countable number of zeroes is zero. The answer is YES if we are dealing with an uncountable infinity,[30] *because probability is not perfectly additive: the sum of an uncountable infinity of zeroes can be positive.*

In the case of the answer N, there are, however, two subcases to be distinguished with reference to question (I) (for which, in cases A and C, the answer can only be YES).

N': Probability zero implies impossibility. What has been said above is a consequence of this identification.

N'': Probability zero does not imply impossibility. However, the behaviour is the same: even if we take the union of them all, the events of probability zero form an event with zero probability.

3.11.3. Let me say at once that the thesis we support here is that of A, *finite additivity*; explicitly, *the probability of a union of incompatible events is greater than or equal to the supremum of the sums of a finite number of them*. Apart from the present author, it would seem that only B.O. Koopman (1940) has systematically adopted and developed this thesis. Others, like Good (1965), admit only finite additivity as an axiom, but do nothing to follow up this observation. Others again, like Dubins and Savage (1965), make use of finite additivity for special purposes and topics.

The thesis N is supported, as far as I know, only by certain logicians, such as Carnap, Shimony and Kemeny (as a consequence of a definition of 'strict coherence').[31]

---

30 I do not know whether this corresponds exactly to the conception of the supporters of this thesis (often one only talks about the case of the continuum).

31 In addition to these serious authors, there is no point in mentioning the large number who refer to zero probability as impossibility, either to simplify matters in elementary treatments, or because of confusion, or because of metaphysical prejudices.

The thesis *C* is the one most commonly accepted at present; it had, if not its origin, its systematization in Kolmogorov's axioms (1933). Its success owes much to the mathematical convenience of making the calculus of probability merely a translation of modern measure theory (we shall say a lot more about this in Chapter 6). No-one has given a real justification of countable additivity (other than just taking it as a 'natural extension' of finite additivity); indeed, many authors do also take into account cases in which it does not hold, but they consider them separately, not as absurd, but nonetheless 'pathological', outside the 'normal' theory.

3.11.4. Let us review, briefly, the main objections to the various theses (we number them: *A*1, *A*2,…; *N*1, *N*2,…; *C*1, *C*2,…). Our point of view is, of course, represented by the objections to *N* and *C*, and by the answers (*A*1*a*, *A*1*b*,…; *A*2*a*, *A*2*b*,…) to the objections raised against *A*. We will also interpolate some examples (*E*1, *E*2,…).

*A*1 This is an objection from the standpoint of *N* (or rather *N*′): *it is not sufficient to exclude as inadmissible those bets with gain X certainly negative* ($\vdash\!\!\!-\!\!\!- X < 0$: *weak coherence*)*; it is necessary to exclude them if the gain is certainly nonpositive* ($\vdash\!\!\!-\!\!\!- X \leqslant 0$: *strict coherence*)*. This means that 'zero probability' is equivalent to 'impossibility'.*

The most decisive reply will be objection *N*2, but it is better not to evade a reply that clarifies the points (perhaps persuasive) put forward in *A*1; this reply will constitute a preliminary refutation of *N* (*N*1).

*A*1*a* It should be unnecessary to point out that the inadmissibility of a bet is always relative to the set of choices offered by a given scheme. It is obvious that if among the possible choices there was the choice 'do not make a bet at all', nobody would choose an alternative that could only lead to losses (this, however, means nothing).

*A*1*b* In the simplest scheme, let $X = -E$ (loss = 1 if *E* occurs; e.g. the risk we are facing), and consider the appropriateness of insuring oneself by paying a premium *p*. Let us suppose that one is willing to pay $\left(\frac{1}{2}\right)^n$ (and no more) if $p = \left(\frac{1}{2}\right)^n$; for example *E* = all heads in *n* tosses. If *E* = all heads in an infinite number of tosses, I will not be willing to pay more than zero (every $\varepsilon > 0$ is $\left(\frac{1}{2}\right)^n$ for sufficiently large *n*, and would be too much even if the risk were infinitely greater). The lesser evil, therefore, is not to insure oneself; in other words, to act in this respect (*but not in others*) as if *E* were impossible.

*A*1*c* There is more, however. The condition of coherence is and must be (as we established in Sections 3.3.5 and 3.3.6) *even weaker*[32] than the one criticized in *A*1, allowing in addition bets in which one can *only lose*! Let us suppose that an individual is subjected to a certain loss of a sum $1/N$ (where *N* is an 'integer chosen at random', with equal – and therefore zero – probabilities for each value, and hence for each finite segment $N \leqslant n$ (II!)). There is no advantage in paying a sum ε (however small) to avoid this certain loss, because it would always be practically certain that the loss avoided would be very much smaller.

*N*1 = *A*1*d* Summarizing and concluding, we have the following. The variants (from the weakest to the strongest) consist in excluding *X* if

$$\sup X < 0, \quad \sup X \leqslant 0, \text{ with } X = 0 \text{ impossible}, \quad \sup X \leqslant 0,$$

Objection *A*1 criticizes the middle statement, and supports the last one. In A1*c* we explained why, on the contrary, we think it necessary to support the first one.

32 If we wished to give this condition a name, we might call it *sufficient coherence* (in contrast to *weak* and *strict* coherence).

*N*2 *The variant N′ is logically absurd unless one excludes the possibility of considering a partition with an uncountable infinity of possible cases* (e.g. the continuum). In the denumerable case objections arise which also apply to *C* (*C*3 = *N*4, and so on).

*N*3 The variant *N″* does away with *N*2: nevertheless, *the meaning of zero probability is still exceptionally restrictive* (much more so than in *C*, and even there it is too restrictive; see *C*4).

In fact, one should be able to define E^* = the union of all events with zero probabilities = the maximal event with zero probability (let us call it 'the catastrophe'). Under the noncatastrophic hypothesis (with probability = 1) one goes back to *N′*; only in the opposite cases are events with zero probability no longer impossible (and, consequently, (II!) can have any probability whatsoever).

3.11.5. *C*1 *C appears to be less logically plausible than A and N* – we suspect 'Adhockery for mathematical convenience' – because the distinction between finite and infinite has without doubt a logical and philosophical relevance, whereas it might seem strange to draw the crucial distinction between finite and nondenumerable on the one hand, and countable on the other hand.

*C*2 A difficulty that derives from this is the following: *given a partition* (*e.g. whose cardinality is that of the continuum*) *into events of zero probability*, what happens if as a consequence of additional information one believes that *only a countable infinity remain possible*? In particular, if one assumes them (II!) equally probable? Or under the most general hypothesis?

*E*1 Initially, *X* has a uniform distribution over the real numbers between 0 and 1 (all points equally probable (II!)). Additional information reveals that *X* is rational.

*E*2 It seems obvious (but recall (II!)) that in this case – that is *E*1 after the given 'additional information' – *the values which remain possible, that is the rational values of* [0, 1], *are* (*still*) *equally probable* (they define a 'random choice' from the original set).

If one thought of actually interpreting the problem geometrically, one might perhaps doubt the judgment of all the rationales as equally probable, considering as 'rather special' the end-points, mid-point, fractions with small denominator, decimal fractions with only a few figures and so on.

This effect is lessened if one thinks of taking the 'distance between two points chosen at random' (the first minus the second; if negative add 1, take the result mod 1).

It disappears altogether if one thinks in terms of a circle obtained by rolling up the segment without indicating which is the 'zero' point.

*C*3 = *N*4 Objection *C*2 can also be raised in the countable case (and then it also concerns *N*). *Suppose that we have a countable infinity of possible cases, one with p* = 1 (*and the others therefore with p* = 0); *assume we know* that *the first one has not occurred*.

*E*3 Let *N* be the number of passages through the origin in a random walk for which $\mathbf{P}(N > n) = 1$ for all *n* (an example is Heads and Tails); information: $N \neq \infty$.[33]

33 This information could only be given by somebody who had explored the world as it appears after the end of time…. Objections to 'lack of realism' would, however, be out of place here as it is merely a question of logical compatibility. Where they are appropriate (and usually insufficiently dealt with), the exigencies of realism will be examined here (especially in the Appendix), perhaps at greater length than hitherto, and perhaps more than is reasonable. One cannot refute the exact nature of a conclusion based on the examination of a 'pathological' curve (e.g. that of Helge van Koch) by the pretext that there exist neither pencils, nor sheets of paper, nor hands, by means of which it could be drawn.

*E*4 In general, in such cases it is plausible to say that the

$$p_h = \mathbf{P}\big(N = h \,|\, N \neq \infty\big)$$

are all zero, and (II!) each is infinitely greater than the preceding one. We limit ourselves to a mere statement of this in order to be able to refer to this example without examining it deeply.

*C*4 *The meaning of p* = 0 *is too restrictive even in C* (although much less so than in *N*; see *N*3). Expressed in a vague form, but one which corresponds exactly to the state of things, this is the 'essence' of those considerations and examples already given (*C*2, *C*3, *E*1, *E*2) and of those to come. The fact that, whereas, for any finite *n*, uniform partitions are allowed (all $p = 1/n$), in the countable case only extremely unbalanced partitions are allowed (under *C* and *N*), may serve as a 'symptom', which makes this 'restrictiveness' appear pathological.

We shall see, on the one hand, just how unbalanced they are, and, on the other hand, the objections to which this gives rise from a realistic point of view. The latter, of course, will vary according to the conception one holds.

*C*5 = *N*5 By taking the sum of probabilities to be = 1 (suppose we denote the probabilities by $p_1, p_2, \ldots, p_i, \ldots$, in decreasing order), one necessarily has an inequality such that for any $\varepsilon > 0$, however small, a finite number of events – the first n_ε – together have probability $>1 - \varepsilon$, and the infinity of the others together have probability $<\varepsilon$. (In such circumstances, I am tempted to say that the events 'are not countably infinite' but 'a finite number – up to trifles').

*E*5 The point made in *C*5 = *N*5 appears even more strange if we take as an example the following observation.

If, instead of the whole infinity of events, one only had the first $N = n/\varepsilon$ (where ε and $n = n_\varepsilon$ are as in the preceding case), there would be nothing to prevent one judging them equally probable (or almost so) in accordance with some assumed reasons or opinions. The total probability of the first *n* would then have been ε instead of $1 - \varepsilon$. Of course, even the infinity of probabilities could have all been taken $<1/N$, but the enormity of the inequality would reappear if we took some $n' = n'_\varepsilon$ and $N' = n'/\varepsilon$ to start with.

From a mathematical standpoint this is obvious What is strange is simply that a formal axiom, instead of being *neutral* with respect to the evaluations (or, for those who believe in them, with respect to the objective reasons), and only imposing formal conditions of coherence, on the contrary, imposes constraints of the above kind without even bothering about examining the possibility of there being a case against doing so.

3.11.6. Let us try to better imagine the reactions of individuals with different points of view.

*C*6 = *N*6 Suppose we are given a countable partition into events E_i, and let us put ourselves into the subjectivistic position. An individual wishes to evaluate the $p_i = \mathbf{P}(E_i)$; he is free to choose them as he pleases, except that, if he wants to be coherent, he must be careful not to inadvertently violate the conditions of coherence.

Someone tells him that in order to be coherent he can choose the p_i in any way he likes, so long as the sum = 1 (it is the same thing as in the finite case, anyway!).

The same thing?!!! You must be joking, the other will answer. In the finite case, this condition allowed me to choose the probabilities to be all equal, or slightly different, or

very different; in short, I could express any opinion whatsoever. Here, on the other hand, the *content* of my judgments enter into the picture: I am allowed to express them only if they are unbalanced to the extent illustrated in *C*5–*N*5–*E*5. Otherwise, even if I think they are equally probable – as I would do in the case of *E*2 – I am obliged to pick 'at random' a convergent series, which, however I choose it, is in absolute contrast to what I think. If not, you call me *incoherent*! In leaving the finite domain, is it I who has ceased to understand anything, or is it you who has gone mad?

*C*7 = *N*7 In the same situation, an objectivist of the classical school finds himself facing case *E*2 (for him 'in conditions of symmetry all possible cases are equally probable').

This much is obvious: the infinite number of cases is equally probable and, therefore, they all have probability $1/\infty = 0$ (perhaps – he may think – I am not expressing myself in an orthodox fashion; the conclusion, however, is this one). To the objection of the teacher who wants a series with sum = 1, and who is not worried if one asks him whether he really wants an opinion so unbalanced as to give rise to the points raised in *E*5, he too will cry out: Is it I who has ceased to understand anything, or is it you who has gone mad?' And he will explain: 'I swear that I find myself in the ideal conditions of complete ignorance, with the absence of any reason to doubt whether any point has objective probability greater than that of any other one. In no other case can I be so sure of being able to state with precision that the objective probabilities are equal, because it is only in this case, where I cannot even see or distinguish the rational points, that I have reached the final sublime peak of total and unsurpassable ignorance. And now, what is the use of it? What are the objective probabilities I must give the various points, and how do I know which of them must be assigned a large probability, a small one, or a very small one?'.

*C*8 = *N*8 For the frequentist, this is even easier. If he thinks of a sequence of experiments (an ideal version of roulette, reduced to a point-ball which can stop at any rational point of the circle of *E*2) he will be in doubt as to whether a point will appear just a few times, or many times, or even infinitely many times. It is unlikely, however, that he will think for a moment that some point – and especially one which can be individuated right from the beginning – will appear so often as to have a limit-frequency different from zero.

*C*9 = *N*9 Here is a new and genuine mathematical objection to countable additivity: for those who conceive of probabilities as limit-frequencies (over a sequence, or, in von Mises' terminology, a 'Collective'), the fact that *limit-frequencies must satisfy finite additivity*, *but not countable additivity*, should be decisive.

(So far as I know, however, none of them has ever taken this observation into account, let alone disputed it; clearly it has been overlooked, although it seems to me I have repeated it on many occasions).

3.11.7. *C*10 A probability which is countably (but not perfectly) additive cannot be defined on the power set of the infinite set of events under consideration.

Therefore, it is necessary:

a) either to introduce restrictions that only allow one to refer to events given by certain 'subsets', excluding the others (in this case the logical justifications are not obvious, and the mathematical ones, which require the creation of special events by endowing the 'space' with topological properties, seem merely to have the status of 'Adhockeries for mathematical convenience');

b) or to accept perfect additivity, that is *N*, which appears *more logical* than *C*, for this reason in addition to that already given in *C*1 (but one encounters *N*2, and abandons any treatment in the continuum, even by means of the measure-theoretic model which is the actual aim of *C*);
c) or to accept finite additivity; that is *A*.

3.11.8. Do there exist objections to *A* (besides *A*1, which we have examined already)? In all honesty – and I shall willingly change my mind if any contrary evidence is brought to my attention – it seems to me that one should in general refer to prejudices and habits, rather than to objections. Independently of the discussion of specific aspects of the real problem (which are always neglected), it is these habits and prejudices which lead one to consider as 'natural', or 'absurd', those things in other branches of mathematics that are more or less customary, more or less up-to-date, and, above all, more or less 'convenient'. We refer to those fields where, in the absence of an intrinsic meaning, already existing and imposed from the outside onto the possible translations into mathematical definitions and axioms, it is admissible to choose those concepts and hypotheses that are most convenient, to choose them 'for mathematical convenience'.

We shall see something of these aspects and attitudes in Chapter 6 and in the Appendix. (It is often difficult to analyse them because they are more psychological than mathematical in character, and because one usually has to deduce things from odd comments rather than from explicit and systematic explanations.) If one wants to pick out an example of a sufficiently concrete position, having some validity,[34] I merely point to the following.

*A*2 It seems to many people that a countable partition that is not unbalanced (i.e. not reducing to cases 'finite up to trifles', as we jokingly called them in *C*5) is 'not feasible'. A positive integer *N*, unknown (random) and capable of taking on any value (between 0 and ∞, which is excluded), is always, in any practically or conceptually imaginable example, almost certainly not too large (and an upper bound is not given solely in order to avoid a more or less arbitrary choice). A partition of a set whose cardinality is that of the continuum, for example an interval, into a countably infinite number of (*L*-) measurable sets, is necessarily such that all the measure (except an arbitrarily small residual) is given by a finite number of them. They can be overlapping (as in the Vitali case) but then they are not measurable and, therefore, not even 'mentionable', and not even susceptible of a constructive description independently of the axiom of choice.

It is necessary to reply to this from various viewpoints.

*A*2*a* From the subjectivistic point of view – since, subject to the conditions of coherence, one has complete freedom of choice in evaluating the probabilities – one can perfectly well assign greater probability to a set with only one point than to a set which has very large measure, or is non-measurable. Conversely, can this line of argument justify attributing large probability to sets consisting of a single point and with small measure, and negligible probability to the large sets, leaving out the intermediate cases?

34 I hope that the reader can himself demolish the frequent attempts to 'prove' countable additivity under the tacit assumption of the validity of some property equivalent to it.

*A*2*b* Do not these examples themselves (although in a slightly more sophisticated manner) reveal the prejudice of assuming the measure-theoretic model as the universal one?

*A*3 Another plausible objection: all these examples and counterexamples are artificial, with no practical interest; there is no reason to prefer a less convenient theory simply because it allows us to take account of them.

*A*3*a* The examples have a critical function; to test the logical consistency of the various points of view. To accept the point of view which (I hope) they reveal to be the logically correct one does not imply that one has to occupy oneself with matters of this nature,[35] but only to avoid expressing oneself in a way that appears to be incorrect (albeit with reference to 'pathological' examples).

*A*3*b* Indeed, in practice, it will probably turn out to be advisable to limit oneself to *even simpler* ideas, sticking to the more elementary ambit (Jordan–Peano measure, Riemann integral) where the conclusions are unexceptionable, rather than passing to the more 'modern' set-up (Borel or Lebesgue measure, Lebesgue integral), given that the usual extension is based on a convention which is inadmissible as a general axiom, and difficult to justify in a realistic way as a particular hypothesis for individual practical cases. It seems to me that it is difficult to justify not only its validity, but even that possible interpretations and applications to actual and practical problems are not illusory.

*A*3*c* If we are going to talk about which theory is 'less convenient', we must distinguish the sense in which 'convenient' is to be understood. The theory given by *C* is, in general, more convenient to handle, and is convenient because it provides a well-determined answer in many cases where *A* just gives bounds. From the standpoint of *A*, it is wrong to substitute an exact answer in place of these bounds (and, anyway, inconvenient, since it forces us to exclude all those examples that might appear artificial, but which are not absurd). From some points of view, *A* is even more tractable; for example every limit of a probability distribution is, in *A*, a probability distribution (possibly not proper): this is not true in *C*. It is, in any case, a question of things which are logically relevant, not one of mathematical convenience.

*A*4 One more objection (a little premature as far as the applications it refers to are concerned, but not in terms of its formal meaning, nor for the understanding of example *E*6 below).

Proofs made in the spirit of *A* in order to invalidate the interpretations of *asymptotic results* (not yet discussed) as *limit-results* (deduced in accordance with concept *C*) often make use of the device of introducing a number *N*, which is 'chosen at random' (zero probability for each single n and finite segment $N \leqslant n$), assuming that from *N* onwards a certain process proceeds in a different way from that foreseen in the scheme of description.

This said, the objection is: *That's a different story: if the scheme changes, if there is a violent change, then the conclusions established under the assumption that the scheme remains unaltered, without foreseeing any possibility of a violent change, will certainly break down.*

---

35 Let us recall that the critical examples which Peano inserted into Genocchi's lecture notes, in order to show that certain 'theorems' did not always hold in 'pathological' cases, met with an exactly similar attitude of disapproval and incomprehension.

*A*4*a* Statements of this kind do not take account of the situation. The 'scheme', as usually described, does not explicitly foresee the possibility of a violent change, but it does not exclude it either: it is entirely neutral. It is, therefore, improper to refer to a 'violent change': the question of a violent change arises only when one adds to the mathematical scheme something more in the way of interpretation, which would be difficult to express. Indeed, if it were expressed, it would render trivial the result, which is beautiful and true only if one assumes that countable additivity is less restrictive than would appear from the following kind of example.

*E*6 As in *E*2, we can imagine 'choosing at random' a rational number in [0, 1] with a finite number of decimal places (all with the same probability (II!)),[36] the number of places being itself random, and not preassigned. If we think of a selection of the successive decimals (or of their successive deciphering or calculation, if they have been 'drawn all at once' and can be worked out successively, as for π), the process is clearly identical to that of drawing any real number whatsoever. At each drawing, all 10 figures have the same probability $\frac{1}{10}$, whatever the previous results may have been.[37]

If by 'catastrophe' we mean the exceeding of the last nonzero figure, it is certain that sooner or later this will happen. *But it will not be a catastrophe*: we will not be able to realize it; nothing will change in the described scheme. Even after 100 or 1 000 000 or 10^{1000} consecutive zeroes, provided we have no gift of divination, the probability that the next figure will be zero is $\frac{1}{10}$, as for any other figure; the probability that the next 100 figures will all be zero is 10^{-100}, as for any other 100-figure number; the probability that the figures will continue to be zero for evermore is zero, exactly as it is at any other instant, and after any arbitrary sequence of figures.

In this example, all the probabilistic assumptions explicitly stated for the process hold exactly; these lead to the conclusion that, with probability = 1, the 10 figures will each show up with limit-frequency $\frac{1}{10}$ (whereas, the limit-frequency is here = 1 for the figure 0, and = 0 for the others). The only assumption that does not hold is that of countable additivity, but if anyone considers it as an axiom, instead of a particular restriction (not valid in our example), he has the right (?) to omit its explicit statement and to check whether it holds.

3.11.9. *Conclusion* (*for the time being*). I do not know whether, and to what extent, the arguments put forward here have been persuasive. On the other hand, it is premature to accept or reject them before encountering other aspects of them and having seen their implications (in Section 3.12 following, at the end of Chapter 4, and in Chapter 6 and elsewhere, more or less incidentally). In view of this, however, I would like to have succeeded in convincing the reader of one thing; that we are dealing with a complex of

---

36 If one wishes, instead of choosing from this set one can imagine the choice of any rational whatsoever, as in *E*2. The rationals can be put together in 'equivalence classes' (where two numbers differ by a bounded decimal fraction; i.e. they coincide from some point on) and in each class *an identifiable representative can be chosen*; the one which is *periodic right from the beginning*. Every rational uniquely determines the components $r = p + d$ (p periodic, d decimal), and the sets I_d (of the r with the same d) give rise to a partition of the rationals into a countable number of sets superposable by translations (mod 1). To choose r is therefore a way of choosing d.
The partition is similar to that of Vitali for the reals, but here, fortunately, an infinite number of choices is not required.

37 We should refer to stochastic independence, but we shall come to it in the next chapter, Chapter 4, and here content ourselves with just mentioning the idea.

problems, connected and meaningful, concerning which there are many things to be discussed under various headings: the conceptual, the mathematical, the practical. It is not just, as might seem logical at first sight, a question of arbitrary conventions for the subtleties involved, having no connection with real problems.

3.12 Random Quantities with an Infinite Number of Possible Values

3.12.1. The above considerations obviously also apply to the case in which there are an infinite number of possible values for a random quantity X. Some new features also arise, however. We shall not concern ourselves with the general case until Chapter 6, but in the meantime it is necessary to mention certain refinements, although only for the more elementary case (elementary in a certain sense, at least) of a countable infinity of possible values x_h ($h = 1, 2, \ldots$). To these will correspond – or rather can be attributed by the person who evaluates them – probabilities p_h, either positive or zero (they might even all be zero), with

$$\sum_h p_h = 1 - p^* \leqslant 1, \left(0 \leqslant p^* \leqslant 1\right).$$

For any interval or set I, one could say, knowing only the x_h and p_h, that $\mathbf{P}(X \in I) = \sum_h p_h(x_h \in I)$ if the set contains a finite number of points, but only that

$$\sum_h p_h\left(x_h \in I\right) \leqslant \mathbf{P}\left(X \in I\right) \leqslant \sum_h p_h\left(x_h \in I\right) + p^*$$

if it contains an infinite number (given that the probability p^* can always be imagined as deriving solely from these).

3.12.2. In particular, if x is an accumulation point of the x_h (it does not matter whether it is one of them or not), we can have nonzero *adherent* probabilities, the latter defined to be the limit of $\mathbf{P}(x - \varepsilon < X < x)$ or $\mathbf{P}(x < X < x + \varepsilon)$ as $\varepsilon \to 0$ ($\varepsilon > 0$), and their sum (if we wish to distinguish, we refer to adherent from the left, adherent from the right). The adherent probabilities (or masses) cannot exceed p^*; not even if we take them all together, or even include those possibly adherent (from the left) to $+\infty$ and (from the right) to $-\infty$.[38] The adherent probabilities could not only have total probability $<p^*$ but also zero (in other words, nonexistent), although p^* was positive, or even $p^* = 1$. As an

38 One can either allow $+\infty$ and $-\infty$ to also appear among the possible values, or one can exclude them. Including them would entail thinking of X as a random point on the completed real line (compactified) with the adjunction of the 'extremes' $+\infty$ and $-\infty$. There is nothing absurd about this, although it is not usual to do and there is no point in insisting upon it. Every now and again we will make brief mention of such eventualities, but without entering into any obligation to observe case by case whether what is said is valid there also.

On the other hand, we must note a certain conflict of interest. As far as prevision is concerned (and here the inequalities are essential), the values $+\infty$ and $-\infty$ are distinct and very far apart (in fact, opposite). From an analytic point of view, however, it would be more natural to consider them as a single value (except for looking at it in terms of approaching from the left and right), thinking, for instance, of the complex sphere (and, in that context, of the circle of real numbers) and of functions which are 'continuous' there, like $y = 1/x$ at $x = 0$ (see *Matematica logico-intuitiva*, 3rd edn, pp. 124–133).

example: X = rational between 0 and 1, with the probability of each subinterval equal to its length (the uniform distribution).

3.12.3. The argument concerning the prevision $\mathbf{P}(X)$ is new and specific to this case. It is unnecessary to note that whatever one says concerning $\mathbf{P}(X)$ holds for any $\mathbf{P}(\gamma(X))$, where $Y = \gamma(X)$ is any function of X, whose possible values are $y_h = \gamma(x_h)$ with probabilities p_h (except that, if one of these values corresponds to an infinite number of the x_h, its probability may be, if $p^* > 0$, greater than the sum of the p_h instead of being equal to it).

What does the knowledge of the possible values x_h and their probabilities p_h allow us to say concerning $\mathbf{P}(X)$? Or rather, expressing ourselves in terms of what the question means in a (subjective) probabilistic sense, what restrictions does the knowledge of the x_h and an existing evaluation of the p_h (which we wish to remain coherent) impose on us when it comes to evaluating the prevision of X?

It is convenient to begin with the case of a *bounded* random quantity X, and to consider directly the minimum and the maximum of the accumulation points, which we denote by x' and x''; we therefore have

$$-\infty < \inf X \leqslant x' \leqslant x'' \leqslant \sup X < +\infty.$$

Let us prove that if $p^* = 0$ (i.e. if $\sum_h p_h = 1$, as it is if countable additivity holds) we must have the unique result $\mathbf{P}(X) = \sum_h p_h x_h$, as in the finite case. Apart from this special case we can only say that

$$\sum_h p_h x_h + p^* x' \leqslant \mathbf{P}(X) \leqslant \sum_h p_h x_h + p^* x''.$$

Thus, if we are not in the above case, $p^* = 0$, $\mathbf{P}(X)$ turns out to be uniquely determined if and only if $x' = x''$; in other words, if the x_h have a unique accumulation point, hence a limit to which they converge.

Proof. For a given $\varepsilon > 0$, take N sufficiently large so that we have

$$\sum_h p_h (h \geqslant N) < \varepsilon,$$

and put $X = X_1 + X_2 + X_3$ with

$$X_1 = X = x_h \text{ if } h < N, \quad \text{and} \quad \text{otherwise} = 0,$$
$$X_2 = X = x_h \text{ if } h \geq N \quad \text{and} \quad x_h < x' - \varepsilon \quad \text{or} \quad x_h > x'' + \varepsilon, \text{and otherwise} = 0,$$
$$X_3 = X = x_h \text{ if } h \geq N \quad \text{and} \quad x' - \varepsilon \leq x_h \leq x'' + \varepsilon, \text{and otherwise} = 0.$$

We have

$$\mathbf{P}(X_1) = \sum_h p_h x_h (h < N) \to \sum_h p_h x_h \left(x_h \text{ bounded!}\right);$$
$$\varepsilon \inf X \leqslant \mathbf{P}(X_2) \leqslant \varepsilon \sup X,$$

because there are at most a finite number of possible values between $\inf X$ and $x' - \varepsilon$, and the same for those between $x'' + \varepsilon$ and $\sup X$, and the total probability of those between them with $h \geqslant N$ is the sum of a finite number of the p_h for which the sum of the series is $<\varepsilon$. Finally, we have

$$p^*(x' - \varepsilon) \leqslant \mathbf{P}(X_3) \leqslant p^*(x'' + \varepsilon).$$

All this holds for every ε and hence, as $\varepsilon \to 0$, one obtains the given bounds.

Remark. It is most instructive and important to observe that these bounds *cannot be improved on*; in other words, it is actually admissible to evaluate $\mathbf{P}(X)$ by giving it any value whatsoever between the two end-points (inclusive). The p^* resulting from infinite zero probabilities (distributed on the possible x_h; it does not matter if these already have positive probabilities $p_h > 0$ or instead have $p_h = 0$) could well be considered as deriving from an infinite number of the x_h converging towards x', or towards x'', and in any intermediate way.

(In addition, one notes that the proof neither presupposes nor establishes countable additivity: it holds here – as it may hold elsewhere – by virtue of additional assumptions implicit in the definition of the particular case.)

3.12.4. We pass from the case of bounded X to that of X *unbounded*. The case of one-sided unboundedness must be considered separately, and we therefore begin with the case of X unbounded from above (obviously, the analysis holds also for the other case); the general case follows as a corollary.

We also suppose that with certainty $X \geqslant 0$ (i.e. $\inf X \geqslant 0$); in the general case it is sufficient to put $X = X_1 - X_2$, $X_1 = 0 \vee X$, $X_2 = |0 \wedge X|$, in order to reduce everything to random quantities which are certainly nonnegative.

Moreover – in order not to complicate the exposition by encountering anew the circumstances already seen in the finite case – we suppose that there do not exist finite accumulation points. We can, therefore, suppose the x_h to be increasing, and tending to $+\infty$ as h tends to infinity.[39]

Under these conditions, putting

$$P_n = \sum_{h=1}^{n} p_h,\ P = \lim P_n,\ p^* = 1 - P,\ S_n = \sum_{h=1}^{n} p_h x_h,$$
$$S = \lim S_n,$$

we have

$$P_n = \mathbf{P}\left(X \leqslant x_n\right),\ \ 1 - P_n = \mathbf{P}\left(X > x_n\right),$$

p^* = the mass adherent from the left at $+\infty$, or placed at $x = +\infty$, or some here, some there;

$$S_n = \mathbf{P}\left\{X\left(X \leqslant x_n\right)\right\},\ S_n + x_n\left(1 - P_n\right) = \mathbf{P}\left(X \wedge x_n\right)$$

(the previsions of X either 'amputated' or 'truncated' at x_n; i.e. replaced, if X exceeds x_n, either by 0 or by x_n, respectively).

Since each 'truncated X' is always $\leqslant X$, we necessarily have

---

39 It is clear that the conclusions of this special case are essentially valid in general if one considers that $X' \leqslant X \leqslant X''$, where we set X' = (the smallest integer $\leqslant X$), $X'' = X + 1$ (and the unit of measurement can be taken as small as we please); X' and X'' are automatically of the type considered (but to pursue this would introduce things which we reserve for the treatment of the continuous case).

If $X_\infty = +\infty$ exists among the possible values, it is not necessary that the finite possible values be unbounded (and not even that they be infinite in number) in order for us to be in the unbounded case.

$$\mathbf{P}(X) \geqslant S_n + x_n(1 - P_n) \text{ for some } n, \text{ and hence}$$
$$\mathbf{P}(X) \geqslant S + x_n(1 - P) = S + x_n p^* \text{ for some } n$$

(because, if we let n increase in S_n and P_n, while keeping x_n fixed, the expression increases, but less than it would if x_n also were allowed to vary, and tends to the given limit).

It necessarily follows straightaway from this that $\mathbf{P}(X) = \infty$ if $S = \infty$ (the series $\sum_h p_h x_h$ diverges), or if $p^* \neq 0$ (there exists a probability placed at, or adherent to, $+\infty$), or both.

In the opposite case, $p^* = 0$ and S finite (the series of the p_h having sum = 1, and the series of the $p_h x_h$ being convergent), admissible evaluations of $\mathbf{P}(X)$ are given by

$$\mathbf{P}(X) = S = \sum_{h=0}^{\infty} p_h x_h, \text{ or any greater value, including } +\infty.$$

This is proved *by continuity* (and in the next section – Section 3.13 – we briefly discuss that property of continuity which we shall make use of here).

First of all, we set $X'_n = X(X \leqslant x_n)$ (X amputated) with $p'_h = p_h$ for $h \leqslant n$, $p'_h = 0$ for $h > n$, and $p'_0 = \sum p_h (h > n) = \mathbf{P}(X'_n = 0)$; as n increases, all the $p'_h = \mathbf{P}(X'_n = h)$ tend to p_h, but $\mathbf{P}(X'_n) = S_n \to S$.

We then set $X''_n = X'_n + a_n(X > n)$, in other words, X''_n (like X'_n) coincides with X if the latter does not exceed x_n, but when it does we replace it with a_n instead of with 0; a_n denotes the first of the x_h for which $x_h p_0 \geqslant n$.[40] The value a_n already gives a contribution $\geqslant n$, hence we certainly have

$$\mathbf{P}(X''_n) \geqslant n \to \infty.$$

We repeat the conclusions in a schematic form:

$$\text{in the case} \begin{cases} p^* > 0 & \mathbf{P}(X) = +\infty; \\ p^* = 0 \begin{cases} S = +\infty & \mathbf{P}(X) = +\infty; \\ S < +\infty & S \leq \mathbf{P}(X) \leq +\infty; \end{cases} \end{cases}$$

3.12.5. If X is unbounded from above and below, $\mathbf{P}(X)$ is completely undetermined. This is obvious straightaway from the fact that we could always have '$\infty - \infty$'; one can obtain this more rigorously by a passage to the limit in the previous cases (suitably balancing the positive and negative terms).

However, one might consider as *special* the evaluation which consists in taking, both for the positive part $0 \vee X$ and for the negative part $0 \wedge X$, the minimum (in absolute value) admissible prevision – denoting it by $\hat{\mathbf{P}}$ – and setting in general

$$\hat{\mathbf{P}}(X) = \hat{\mathbf{P}}(0 \vee X) - \hat{\mathbf{P}}(|0 \wedge X|) \left(\text{or, briefly, } \hat{S} = S^+ + S^-\right).$$

---

40 The argument, with a simple modification, also holds in the case in which $p_h = 0$ for all possible x_h from a certain $h = N$ on, so that $p_0 = 0$. One could, for instance, let $p_0 = (\frac{1}{2})^n$ taking this probability away from one or more of the p_h (for instance, from p_1 if $p_1 = 0$, starting from that n for which $(\frac{1}{2})^n < p_1$).

'Special' is *not used in a general sense* but if, and so long as, one can consider that, in a given case, the unbounded X is a theoretical schematization substituted for simplicity in place of an actual X, which is in reality bounded, but whose bounds are very large and imprecisely known.

This asymptotic prevision (as we shall call it for this reason) turns out to be:

$$\hat{S} = S^+ + S^- \begin{cases} \textit{finite}, \text{ if } S^+ \text{ and } S^- \text{are;} \\ \textit{infinite}, \text{ if one of the components is : } +\infty \text{ if } S^+ = +\infty; \\ \quad -\infty \text{ if } S^- = -\infty; \\ \textit{undefined}, \text{ if both are infinite.} \end{cases}$$

3.13 The Continuity Property

The property says (and we shall make this precise and prove it) that *coherence is preserved in a passage to the limit*. The property does not hold (without further conditions) when we impose countable additivity. This turns out to be very useful as a tool in proofs of admissibility like the ones just given above (Section 3.12.4).

Theorem. Let $\mathbf{P}_n(E)$ be the evaluations of (coherent) probabilities defined over the same field of events $\mathscr{E}$ (or over different fields of events having $\mathscr{E}$ in common), and put $\mathbf{P}(E) = \lim \mathbf{P}_n(E)$ when it exists (letting $\mathscr{E}' \subseteq \mathscr{E}$ be the set of the E for which the limit exists). In this field the $\mathbf{P}(E)$ itself constitutes a (coherent) evaluation of probability.

Remark. In place of the (more 'familiar') formulation above, it would be (mathematically) preferable to substitute that in which one speaks of the prevision of random quantities rather than the probability of events, and hence of linear spaces (with appropriate definitions and convergence) rather than 'fields'.

Proof. The conditions of coherence are expressed by linear equations (or inequalities) involving a finite number of elements (events, or random quantities); in the passage to the limit these are preserved.

Remark. In a more expressive formulation (and more precise, so long as one recalls that the meaning of 'convergence' is that given above): an evaluation of probability $\mathbf{P}$ adhering to a set $\mathscr{P}$ of coherent evaluations is coherent.

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*, $\mathbf{P}(E|H)$ – and, more generally, of *conditional prevision*, $\mathbf{P}(X|H)$ – 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 $\mathbf{P}(HE)$, I can be sure of receiving one lira if HE occurs; but I can obtain the same result by paying $\mathbf{P}(E|H)$ only if I know H is true, and I can arrange for this amount, $S = \mathbf{P}(E|H)$, in the case of the occurrence of H by paying S. $\mathbf{P}(H)$ now; hence

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

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 $H = 0 = \textit{false}$; $L = \left[(X - \bar{x})/k\right]^2$ if $H = 1 = \textit{true}$.)

$\mathbf{P}(X|H)$, *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 $\mathbf{P}(E|H)$, 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 $\mathbf{P}(X|H)$, $\mathbf{P}(H)$ and $\mathbf{P}(HX)$, 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

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

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

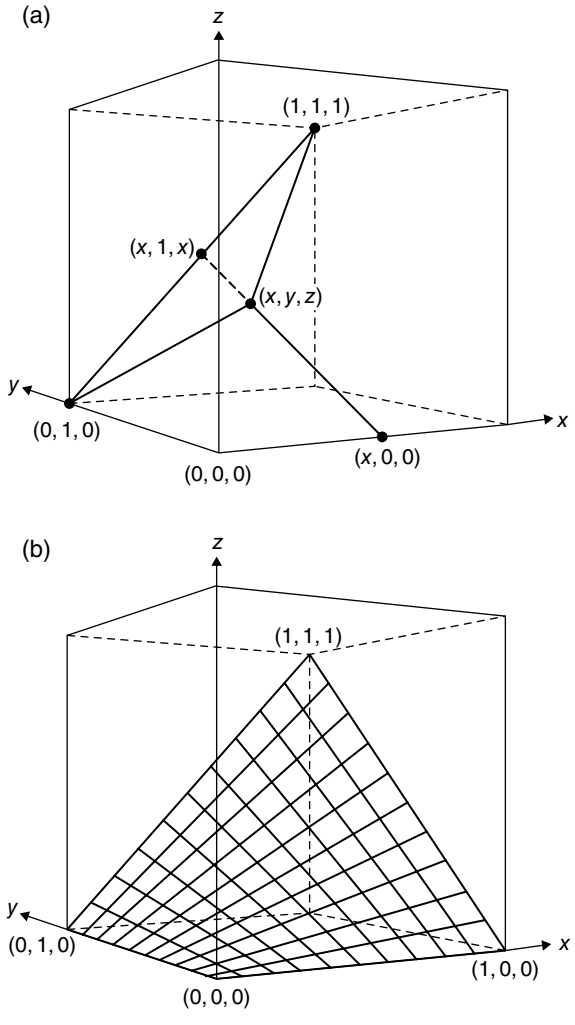


Figure 4.1 The two diagrams illustrate, in two stages, the argument given in Section 4.3: (a) shows why the prevision-point (x, y, z) must lie on a generator of the paraboloid $z = xy$ (presenting visually the argument of the text); (b) shows the set of all possible prevision-points (the part of the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

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

If (x, y, z) were not on the paraboloid $z = xy$ (i.e. not in the plane through the line $y = 1$, $z = x$ 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

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

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 $(x, 1, x)$, 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 $1 = \textit{true}$, $0 = \textit{false}$, $\emptyset = \textit{void}$, 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:

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

which implies that

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

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

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

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

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

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 $\mathbf{P}$ (an element of P in the linear ambit we started with). In fact,

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

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

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

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

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

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:

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

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,..., 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 $\mathscr{H} = (H_1, H_2, \ldots, H_S)$, and the probabilities, $\mathbf{P}(E|H_j)$, of an arbitrary event E conditional on each of the H_j. Since $EH_1 + EH_2 + \ldots + EH_S = E(H_1 + H_2 + \ldots + H_s) = E.1 = E$, and $\mathbf{P}(EH_j) = \mathbf{P}(H_j)\mathbf{P}(E|H_j)$, one has

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

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:

$$\min \mathbf{P}(E \mid H_j) \leq \mathbf{P}(E) \leq \max \mathbf{P}(E \mid H_j) \tag{4.7}$$

(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,

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

we can write the expression above as

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

More generally, we have, of course,

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

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 $\mathbf{P}(EH)$; in fact,

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

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 ~), *the probability is expressible in terms of the* p_i *by the same formula*.[7] For example, if

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

expanding, we obtain

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

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

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

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

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

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]

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

---

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_i as must be the case by virtue of the idempotence of the E_i, E_i^k. For example, if $E = \left(E_1 \vee E_2\right)\left(E_1 \vee E_3\right) = \left(E_1 + E_2 - E_1E_2\right)\left(E_1 + E_3 - E_1E_3\right)$ and we substituted straightaway, we would wrongly obtain $\mathbf{P}\left(E\right) = \left(p_1 + p_2 - p_1p_2\right)\left(p_1 + p_3 - p_1p_3\right) = p_1^2\tilde{p}_2\tilde{p}_3 + p_i\left(\tilde{p}_2p_3 + p_2p_3\right) + p_2p_3$, 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 $\mathbf{P}(E) = \left(\frac{1}{2}\right)^n \times$ 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

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

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 ith 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

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

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 $\mathbf{P}(E) = \mathbf{P}\left(E_i'\right)\mathbf{P}\left(E_i''\right)\mathbf{P}\left(E_h'''\right)$ 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 … 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

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

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 $H_1 \ldots H_s$ instead of two, with probabilities $c_1 \ldots c_s$, and with

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

gives:

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

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:

$$\mathbf{P}(E \mid H) = \mathbf{P}(E \mid H_1 H_2 \ldots H_n) = K\mathbf{P}(E)\mathbf{P}(H_1 \mid E)\mathbf{P}(H_2 \mid E)\ldots\mathbf{P}(H_n \mid E). \qquad (4.19)$$

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

$$\frac{\mathbf{P}(E_h \mid H)}{\mathbf{P}(E_k \mid H)} = \frac{\mathbf{P}(E_h)}{\mathbf{P}(E_h)} \cdot \frac{\mathbf{P}(H_1 \mid E_h)}{\mathbf{P}(H_1 \mid E_k)} \cdot \frac{\mathbf{P}(H_2 \mid E_h)}{\mathbf{P}(H_2 \mid E_k)} \cdots \frac{\mathbf{P}(H_n \mid E_h)}{\mathbf{P}(H_n \mid E_k)}. \qquad (4.19')$$

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,..., 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 α and β 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

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

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 $X_i - \mathbf{P}(X_i)$. 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:

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

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

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.

$$\mathbf{P}(X - x)^2 = \mathbf{P}\left[(X - m) - (x - m)\right]^2$$
$$= \mathbf{P}(X - m)^2 + (x - m)^2 - \{2(x - m)\mathbf{P}(X - m)\},$$

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.

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

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

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

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

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

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

where σ_{ij}, so defined, is called the *covariance* of X_i and X_j,[20] and, writing $\sigma_{ij} = \sigma_i \sigma_j r_{ij}$, 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 σ 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 $\mathbf{P}_Q(X) = \sqrt{(m^2 + \sigma^2)}$, 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. $X \doteqdot Y$). 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 $\boldsymbol{\sigma}(X)\sqrt{(1 - r^2)}$ (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

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

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

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

with a standard procedure (similar to the above), given any linearly independent X_1, X_2, …, 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,…, n), it suffices to add to X_{i+1} a suitable linear combination of X_1,…, 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

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

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 σ_{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/σ_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

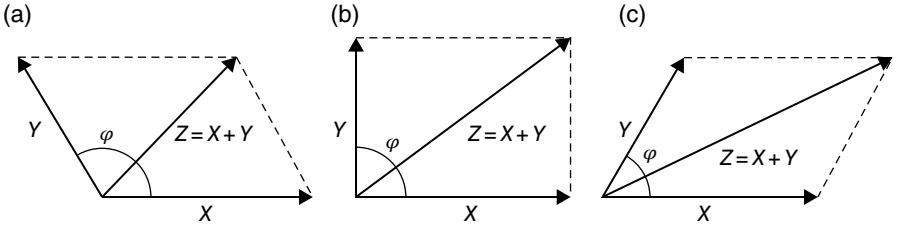


Figure 4.2 (a) Negative correlation. (b) Noncorrelation (orthogonality). (c) Positive correlation.

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

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, $\doteq 0$, 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 $(n-1)$-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 $1/(n-1)$ (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.

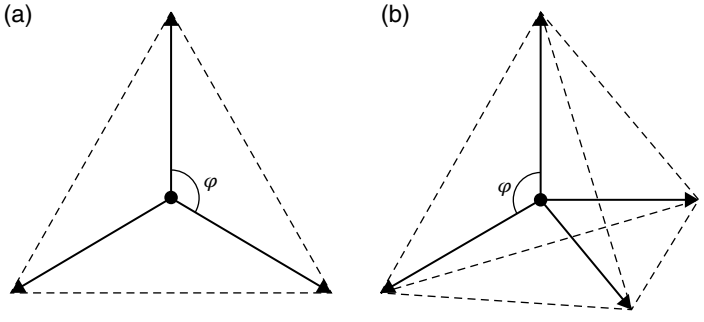


Figure 4.3 (a) The maximum negative correlation for three vectors: $r = \cos\phi = -\frac{1}{2}$. (b) The maximum negative correlation for four vectors: $r = \cos\phi = -\frac{1}{3}$.

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, σ_h, in the ellipsoid of representation, whereas, for the ellipsoid of covariance, they give the reciprocals, $1/\sigma_h$ (or K/σ_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

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

but it is also r times the moment about a diametrical hyperplane, and hence the latter is $1/(r + 2)$. 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

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

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 α' of a set $\boldsymbol{I}$ to be that for which $\boldsymbol{V}(I_\rho) \sim m\rho^{3-\alpha}$ (I_ρ = 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 ≥ 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 ($\sum_h p_h > 0$) 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 $\liminf \mathbf{P}(E|N \leqslant n) \leqslant \mathbf{P}(E) \leqslant \limsup \mathbf{P}(E|N \leqslant n)$.

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 $\phi(n) = n^{-\alpha}$ $(\alpha > 0)$, we obtain, for example, an event E_α, and each E_α has infinitely greater probability than those with a larger α (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

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

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

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

we have

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

and hence $\mathbf{P}(E) = p$ because $\mathbf{P}\left(H_n^*\right)$ and, *a fortiori*, $\mathbf{P}\left(EH_n^*\right)$ 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,… 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').

5

The Evaluation of Probabilities

5.1 How should Probabilities be Evaluated?

In order to say something about this subject without running the risk of being misunderstood, it is first of all necessary to rule out the extreme dilemma that a mathematical treatment often poses: that of either saying everything, or of saying nothing. As far as the evaluation of probabilities is concerned, one would be unable to avoid the dilemma of either imposing an unequivocal criterion, or, in the absence of such a criterion, of admitting that nothing really makes sense because everything is completely arbitrary.

Our approach, in what follows, is entirely different. We shall present certain of the kinds of considerations that do often assist people in the evaluation of their probabilities, and might frequently be of use to You as well. On occasion, these lead to evaluations that are generally accepted: You will then be in a position to weigh up the reasons behind this and to decide whether they appear to You as applicable, to a greater or lesser extent, to the cases which You have in mind, and more or less acceptable as bases for your own opinions. On other occasions, they will be vaguer in character, but nonetheless instructive. However, You may want to choose your own evaluations. You are completely *free* in this respect and it is entirely your own *responsibility*; but You should beware of superficiality. The danger is twofold: on the one hand, You may think that the choice, being subjective, and therefore arbitrary, does not require too much of an effort in pinpointing one particular value rather than a different one; on the other hand, it might be thought that no mental effort is required, as it can be avoided by the mechanical application of some standardized procedure.

5.2 Bets and Odds

5.2.1. One activity which frequently involves the numerical evaluation of probabilities is that of betting. The motivation behind this latter activity is not usually very serious-minded or praiseworthy, but this is no concern of ours here. We should mention, however, that such motivations (love of gambling, the impulse to bet on the desired outcome, etc.) may to some extent distort the evaluations. On the other hand, motives of a different kind lead to similar effects in the case of insurance, where the first objection does not apply.

However, with all due reservation, it is worthwhile starting off with the case of betting, since it leads to simple and useful insights.

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

5.2.2. An important aspect of the question (one to which we shall frequently return) is the necessity of 'getting a feeling' for numerical values. Many people if asked how long it takes to get to some given place would either reply 'five minutes' or 'an hour', depending on whether the place is relatively near, or relatively far away: intermediate values are ignored. Another example arises when people are unfamiliar with a given numerical scale: a doctor, although able to judge whether a sick man has a high temperature or not, simply by touching him, would be in trouble if he had to express that temperature on a scale not familiar to him (Fahrenheit when he is used to Centigrade, or vice versa). Likewise, in probability judgments, there are also those who ignore intermediate possibilities and pronounce 'almost impossible' to everything that to them does not appear 'almost certain'. If neither YES nor NO appears sufficiently certain to them, they simply add 'fifty-fifty' or some similar expression. In order to get rid of such gaps in our mental processes it is necessary to be fully aware of this and to get accustomed to an alternative way of thinking.

In this respect, betting certainly provides useful experience. In order to state the conditions for a bet, which have to be precise, it is necessary to have a sufficiently sensitive feeling for the correspondence between a 'numerical evaluation' and 'awareness' of a *degree of belief*. In becoming familiar with judging whether it is fair to pay 10, 45, 64 or 97 lire in order to receive 100 lire if a given event occurs, You will acquire a 'feeling' for what 10%, 45%, 64% or 97% probabilities are. Together with this comes an ability to estimate small differences and a sharpening of that 'feeling for numerical values', which must be improved for the purpose, of course, of analysing actual situations.

5.2.3. These two aspects come together in the particularly delicate question of evaluating very small probabilities (and, complementarily, those very close to 1). Approximations that are adequate (according to the circumstances and purposes involved) in the vicinity of $p = \frac{1}{2}$ (e.g. 50% ±5%, ± 1%, ±0.1%) are different from those required in the case of very small probabilities: here, the problem concerns the order of magnitude (whether, for example, a small probability is of the order of 10^{-3}, or 10^{-7}, or 10^{-20},...). In this connection, it is convenient to recall Borel's suggestion of calling 'practically impossible', with reference to *'human*, *earthly*, *cosmic* and *universal* scales', respectively, events whose probabilities have the orders of magnitude of 10^{-6}, 10^{-15}, 10^{-50} and 10^{-1000}. This is instructive if one wishes to give an idea of how small such numbers (and therefore such probabilities) are, provided that no confusion (in words or, worse, in concepts) with 'impossibility' arises.[1]

5.2.4. *On the use of 'odds'.* In the jargon used by gamblers, the usual way of expressing numerical evaluations is somewhat different, although, of course, equivalent. Instead of referring to the *probability p*, which (in the sense we have given) is the amount of a bet, we refer to the *odds*,

1 Borel himself, and other capable writers, fail to avoid this misrepresentation when they give the status of a principle – 'Cournot's principle' – to the confusion (or the attempt at a forced identification) between 'small probabilities', which, by convention, could be termed 'almost impossibility', and 'impossibility' in the true sense. What is overlooked here is that 'prevision' is not 'prediction'. The topic is dealt with in E. Borel, *Valeur pratique et philosophie des probabilités* (p. 4 and note IV), part of the great *Traité du Calcul des Probabilités* which he edited; Gauthier–Villars, Paris (1924) (and subsequent editions).

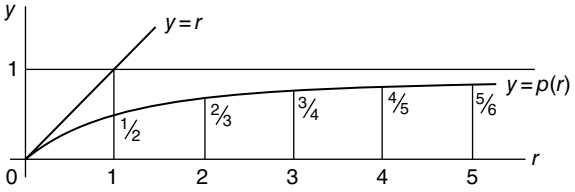


Figure 5.1 The relation between probability (p) and odds (r): $r = p/(1-p)$.

$$r = p/(1-p) = p/\tilde{p}.$$

These are usually expressed as a fraction or ratio, $r = h/k = h{:}k$ (h and k integers, preferably small), by saying that the odds are '*h to k on*' the event, or '*k to h against*' the event. Of course, given r, that is the odds, or, as we shall say, the *probability ratio*, the probability can immediately be obtained by

$$p = r/(r+1), \quad \text{i.e.} \quad (\text{if } r \text{ is written as } h/k) \quad p = h/(h+k). \tag{5.1}$$

A few examples of the correspondence between probabilities and probability ratios, and vice versa, are shown below and illustrated in Figure 5.1:

p	$p/\tilde{p}$	$= r$	$= h/k$	in words	(check) $h/(h+k) = p$
0.20	20/80	=0.25	=1/4	'4 to 1 against'	$1/(1+4) = 0.20$
2/7 = 0.286	28.6/71.4	=0.40	=2/5	'5 to 2 against'	$2/(2+5) = 0.286$
0.50	50/50	=1	=1/1	'evens'	$1/(1+1) = 0.50$
0.75	75/25	=3	=3/1	'3 to 1 on'	$3(3+1) = 0.75$

Observe that to the complementary probability, $\tilde{p} = 1 - p$, there corresponds the reciprocal ratio, $\tilde{p}/p = 1/(p/\tilde{p}) = 1/r$ (i.e. to 'h to k on' there corresponds the symmetrical phrase 'k to h on').[2]

5.2.5. *Extensions*. Probability is preferable by far as a numerical measure (additivity is an invaluable property for any quantity to possess!).[3] However, there are cases in which it is advisable to employ the probability ratio (especially in cases involving likelihood – Chapter 4 – which are often considered in the form of 'Likelihood Ratio') and it

2 It would perhaps be better to introduce a notation to indicate that we are passing from probability to 'odds'; similar to that used for 'complementation' ($\tilde{P} = 1 - p$). An analogous approach would be to take $\underset{\sim}{p} = p/\tilde{p}$ (and if $p = \mathbf{P}(E)$ to use therefore $\underset{\sim}{\mathbf{P}}(E) = \mathbf{P}(E)/\mathrm{P}(\tilde{E}) = \mathbf{P}(E)/\sim \mathbf{P}(E)$). We prefer merely to draw attention to the possibility without introducing and experimenting with more new ideas than prove to be absolutely necessary. To avoid any difficulties, or risks of confusion in notation, we denote the odds more clearly by writing $O(x) = x/(1-x)$, $O[\mathbf{P}(E)] = \mathbf{P}(E)/\mathbf{P}(\tilde{E})$.

3 A newspaper, in considering three candidates for the American presidential election, attributed odds of 2 to 1 on, 3 to 1 against and 5 to 1 against; these are equivalent to probabilities of $\frac{2}{3}, \frac{1}{4}, \frac{1}{6}$, with sum $(8+3+2)/12 = 13/12 > 1$. It is difficult for a slip of this nature to pass unnoticed when expressed in terms of probabilities; using percentages especially, it would certainly not escape notice that 67% + 25% + 17% = 109% was inadmissible.

is useful to indicate, at this juncture, the way in which we shall generalize its use (or, in a certain sense, substitute for it) in cases where the need arises.

In accordance with, and in addition to, the conventions introduced in Chapter 3, Section 3.5, concerning the use of the symbol **P**, we can denote that $r = h/k$ by writing

$$\begin{aligned}\mathbf{P}\left(E,\ \tilde{E}\right) &= \left(h/\left(h\ +\ k\right),\ k/\left(h\ +\ k\right)\right) \\ &= \left(h,k\right)/\left(h\ +\ k\right) = \ K\left(h,k\right) = \left(h:k\right) = \mathbf{P}\left(E:\tilde{E}\right),\end{aligned} \tag{5.2}$$

where we have successively and implicitly made the following conventions:

a common factor, such as $1/(h+k)$, can be taken outside the parentheses; that is $m(a, b) = (ma, mb)$;

such a factor may be taken as understood, denoting it by K, to simply mean that proportionality holds;

the same thing may be indicated by simply using the 'colon' (:) as the dividing sign, rather than the comma. This means that two n-tuples of numbers $(a_1, a_2,\ldots, a_n)$ and $(b_1, b_2,\ldots, b_n)$, not all zero, are said to be proportional if $b_i = Ka_i$ where K is a nonzero constant. Proportionality is sometimes denoted by the sign $\propto$ (which is not very good), and can also be expressed by $= K$. We make the convention – once and for all – that K denotes a *generic* coefficient of proportionality, whose value is not necessarily the same, not even for the duration of a given calculation: we can write, for example, $(2,1,3) = K(4,2, 6) = K(6, 3,9)$. The equals sign is sufficient on its own if the n-tuple with ':' in place of ',' is interpreted as 'up to a coefficient of proportionality' (like homogeneous coordinates); that is as a multi-ratio. Hence, for example, (2:1:3) = (4:2:6) = (6:3:9).

Sometimes the omission of the proportionality factor is irrelevant because it is determined by normalization: for example, if it is known that $E_1\ldots E_n$ constitute a partition, and we write

$$\mathbf{P}\left(E_1 : E_2 :\ldots: E_n\right) = \left(m_1 : m_2 :\ldots: m_n\right), \tag{5.3}$$

it is clear that $\mathbf{P}(E_i) = m_i/m,\ m = m_1 + m_2 + \ldots + m_n$, because the sum must equal 1. In other cases (for any E_i whatsoever, even if they are compatible), one can make the common divisor m enter in explicitly, for example by adding in 1 = the certain event:

$$\mathbf{P}\left(E_1 : E_2 :\ldots: E_n : 1\right) = \left(m_1 : m_2 :\ldots: m_n : m\right). \tag{5.4}$$

The resulting convenience is most obvious when the m_i are small integers. For example, if A, B, C form a partition $(A + B + C = 1)$, by writing $\mathbf{P}(A{:}B{:}C) = (1{:}5{:}2)$(even without the refinement $\mathbf{P}(A{:}B{:}C{:}1) = (1{:}5{:}2{:}8)$) it becomes obvious that

$$\mathbf{P}\left(A\right) = 1/8 = 12\cdot 5\%,\quad \mathbf{P}\left(B\right) = 5/8 = 62\cdot 5\%,\ \mathbf{P}\left(C\right) = 2/8 = 25\%.$$

At this point we shall also introduce the operation of the *term-by-term* product of multiratios, denoting it by *:

$$\left(a_1 : a_2 :\ldots: a_n\right) * \left(b_1 : b_2 :\ldots: b_n\right) = \left(a_1 b_1 : a_2 b_2 :\ldots: a_n b_n\right). \tag{5.5}$$

This frequently provides some advantage in handling small numbers or simple expressions in a long series of calculations, and will turn out to be particularly useful for the applications to likelihood which we mentioned above.

The time has now come to end this digression concerning methods of numerically denoting probabilities and to return to questions of substance.

5.3 How to Think about Things

5.3.1. In discussing the central features of the analysis which must underlie each evaluation, it will be necessary to go over many things which, although obvious, cannot be left out, and to add a few other points concerning the calculus of probability.

The following recommendations are obvious, but not superfluous:

- to think about every aspect of the problem;
- to try to imagine how things might go, or, if it is a question of the past, how they might have gone (one must not be content with a single possibility, however plausible and well thought out, since this would involve us in a *prediction*: instead, one should encompass all conceivable possibilities, and also take into account that some might have escaped attention);
- to identify those elements which, compared with others, might clarify or obscure certain issues;
- to enlarge one's view by comparing a given situation with others, of a more or less similar nature, already encountered;
- to attempt to discover the possible reasons lying behind those evaluations of other people with which, to a greater or lesser extent, we are familiar, and then to decide whether or not to take them into account. And so on.

In particular, in those cases where bets are made in public (e.g. horse races, boxing matches – in some countries even presidential elections) some sort of 'average public opinion' is known by virtue of the existing odds. More precisely, this 'average opinion' is that which establishes a certain 'marginal balance' in the demand for bets on the various alternatives. This might be taken into consideration in order to judge, after due consideration, whether we wish to adopt it, or to depart from it, and if so in which direction and by how much.

5.3.2. In order to provide something by way of an example, let us consider a tennis match between two champions, A and B.[4] You will cast your mind back to previous matches between them (if any); or You will recall matches they had with common opponents (either recently, or a long time ago, under similar or different conditions); You will consider their respective qualities (accuracy, speed, skill, strength, fighting spirit, temper, nerves, style, etc.) and the variation in these since the last occasion of direct or indirect comparison; You will compare their state of health and present form, and so on; You will try to imagine how each quality of the one might affect, favourably or otherwise, his opponent's capacity to settle into the game, to fight back when behind, to avoid losing heart, and so on. For instance, you may think that B, although on the whole a better player, will lose, because he will soon become demoralized as a result of A's deadly service. However, it would be naïve to stop after this first and lone supposition: it would mean to aspire to making a prediction rather than a prevision. You will go on next to think of what might happen if this initial difficulty for B does not materialize,

4 This example has already been discussed by Borel and again by Darmois (see p. 93 of the Borel work mentioned previously, and again on p. 165 Darmois' note VI). As is clear from this and other examples (like his discussion, again on p. 93, of the evaluation of a weight – similar to our example in Chapter 3, Section 3.9.7), Borel seems to be inspired – in the greater part of his writings – by the subjectivistic concept of probability: he can thus be regarded as one of the great pioneers, although incompatible statements and interpretations crop up here and there, as was pointed out in the footnote to Section 5.2.3.

or is overcome, and little by little You will obtain a summary view – but not a one-sided or unbalanced one – of the situation as a whole. Your ideas about the values to attribute to the winning probabilities for A and B will in this way become more precise. You may have the opportunity to compare your ideas and previsions with those of other people (in whose competence and information You have a greater or lesser confidence, and whom You may possibly judge to be more or less optimistic about their favourite). In the light of all this, You might think over your own point of view and possibly modify it.

5.3.3. Our additional remarks concerning the calculus of probability consist in pointing out that the conditions of coherence, even if they impose no limits on the freedom of evaluation of any probability, do in practice very much limit the possibility of 'extreme' evaluations. More precisely, an *isolated* eccentric evaluation turns out to be impossible (the same thing happens, for instance, to a liar, who, in order to back up a lie, has to make up a whole series of them; or to a planner, who must modify his entire plan if one element is altered).

It is easy to say 'in my opinion, the probability of E is, roughly speaking, twice what the others think it is.' However, if You say this, I might ask 'what then do You consider the probabilities of A, B, C to be?', and, after You answer, I may say 'so do You think the probability of H is as small as this; $\frac{1}{10}$ of what is generally accepted?', and so on. If You remain secure in your coherent view, You will have a complete and coherent opinion that others may consider 'eccentric' (with as much justification as You would have in calling the common view 'eccentric') but will not otherwise find defective. However, it will more often happen that as soon as You face the problem squarely, in all its complexity and interconnections, You come to find yourself in disagreement *not only with the others but also with yourself*, by virtue of your eccentric initial evaluation.

We have been talking in terms of bets and the evaluations of probability, and not of previsions of random quantities, although they are the same thing in our approach. This was simply a question of the convenience of fixing ideas in the case where the probabilistic aspect is most easily isolated: however, one should note that the same considerations could in fact be extended to the general situation.

5.4 The Approach Through Losses

The betting set-up is related to the 'first criterion' of Chapter 3, Section 3.3; the scheme we are now going to discuss is based on the 'second criterion': it is this latter – as we remarked previously, and will shortly see – which turns out to be the more suitable.

First of all, we shall find it convenient to present this scheme right from the beginning again, referring ourselves now to the case of events. Because this is the simplest case, and because we are treading an already familiar path – which we shall illustrate clearly with diagrams – everything should appear both more straightforward and of wider application.

5.4.1. Instead of some general random quantity X, You must now think in terms of an event E, such that You are free to choose a value x, bearing in mind that You face a loss

$$L = L_x = \left(E - x\right)^2. \tag{5.6}$$

Expanding this (remembering that $E^2 = E$), one obtains the following alternative expressions (in the last one, p is any number whatsoever):

$$\begin{aligned} &(\mathrm{a}) \quad L_x = x^2 + (1-2x)E, \\ &(\mathrm{b}) \quad\quad = x^2\tilde{E} + (1-x)^2 E, \\ &(\mathrm{c}) \quad\quad = E(1-p) + (p-x)^2 + (E-p)(p-2x). \end{aligned} \tag{5.7}$$

They all reveal (5.7b most explicitly) that L_x equals x^2 or $(1-x)^2$ according to whether $E = 0$ = false or $E = 1$ = true.

Since we have already used the criterion as a definition – and hence already know what the probability $p = \mathbf{P}(E)$ of E is – we can, 'being wise after the event', examine how the criterion behaves by looking at $\mathbf{P}(L_x)$, considered as a function of a value x and of a probability p, assumed to be arbitrary (so we adopt the notation $L_x(p)$). Putting $E = p$ in equations 5.7a, 5.7b, and 5.c (which are linear in E) we obtain:

$$\begin{aligned} &(\mathrm{a}) \quad L_x(p) = x^2 + (1-2x)p \\ &(\mathrm{b}) \quad\quad = x^2\tilde{p} + \tilde{x}^2 p, \\ &(\mathrm{c}) \quad\quad = p(1-p) + (p-x)^2 = p\tilde{p} + (p-x)^2. \end{aligned} \tag{5.8}$$

5.4.2. We now examine the variation in $L_x(p)$ as p varies, x being an arbitrary fixed value. As might have been expected (5.8a shows this up most clearly), L_x varies linearly from $L_x(0) = x^2$ to $L_x(1) = \tilde{x}^2$ (which are the two possible values for L_x, depending on the occurrence of either $\tilde{E}(p = 0)$ or $E(p = 1)$). The straight lines in Figure 5.2, connecting these extreme values, give a visual impression of how they go together: that is of how, in order to reduce the penalty resulting in one case, one must increase it in the other.[5]

The figure also shows, in an indirect way, the variation of $L_x(p)$ for varying x, with p fixed. Geometrically, one can see (and equation 5.8c) presents it explicitly) that the straight lines are the tangents to the parabola $y = p(1-p) = p\tilde{p}$, and that none of them can go beneath their envelope (this is within the interval [0, 1]: the others would correspond to values $x < 0$ and $x > 1$; see footnote 5). Given p, the best one can do is to take the tangent at p, obtained (as we already know!) by choosing $x = p$: this gives $L_x(p)$ its minimum value (as x varies), $L_x(p) = p\tilde{p}$. Choosing a different x gives rise, in prevision, to an additional loss $(x-p)^2$; that is the square of the distance from x to p: equation 5.8c shows this explicitly, by splitting the linear function $L_x(p)$ into the sum of $p(1-p)$ (the parabola) and $(x-p)^2$ (the deviation from the parabola of the tangent at $p = x$). We observe also, and this confirms what has been said already, that this deviation is the same for all the tangents (starting, of course, from their respective points of contact).

The maximum loss is 1, and this is achieved by attributing probability zero to the case that actually occurs: the minimum loss is 0, and is achieved when a probability of 1 (or 100%) is attributed to this case. For any given x, the loss varies between x^2 and $\tilde{x}^2$ (as we have seen already). For a given p, we already know that the minimum is $p\tilde{p}$ (for $x = p$), and it is readily seen that the minimum is $p \vee \tilde{p}$: more precisely, if $p \leqslant \frac{1}{2}$ it is $1-p$, obtained by choosing $x = 1$; if $p \geqslant \frac{1}{2}$ it is p, obtained by choosing $x = 0$. If $p = \frac{1}{2}$, we have the maximum of

5 This is for $0 \leqslant x \leqslant 1$: we already know, and can also see, that, in every case, $x < 0$ or $x > 1$ is worse than $x = 0$ or $x = 1$, respectively, and is thus automatically ruled out (without need of any convention).

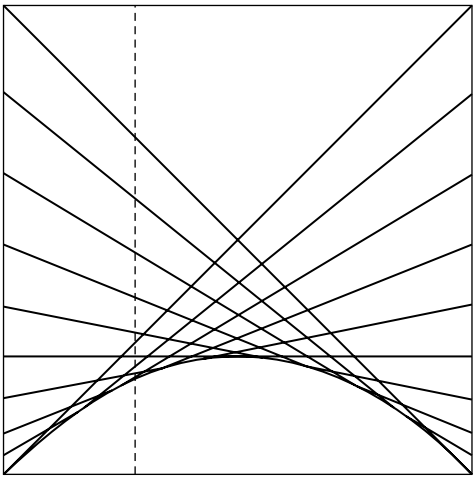


Figure 5.2 The straight lines correspond to the combinations of penalties among which the method allows a choice (the penalty can be reduced in one of the two cases at the expense of increasing it in the other: lowering the ordinate at one end raises it at the other). The ordinate of a particular straight line at the point p is the prevision of the loss for the person who chooses that line and attributes probability p to the event under consideration. In this case, the minimum value that can be attained is given by the ordinate of the parabola (no straight line passes beneath it!), and the optimal choice of straight line is the tangent to the parabola at the point with abscissa p.

the minimum ($p\tilde{p} = \frac{1}{4}$), and the minimum of the maximum ($p \vee \tilde{p} = \frac{1}{2}$), and hence the largest discrepancy ($\max - \min = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$); in general, the discrepancy is $x^2 \vee \tilde{x}^2$, that is the maximum of x^2 and $(1-x)^2$, and attains its maximum (=1) for $x = 0$ and $x = 1$.[6]

5.4.3. *The case of many alternatives*. We can deal with the case of many alternatives (of a multi-event, of a partition), and the more general case of any number of arbitrary but not incompatible events, by applying the previous scheme to each event separately. In this way, things reduce to the treatment given in Chapter 3, and to the geometric representation which was there illustrated. Here, we simply wish to review the approach in the spirit of the above considerations, and then to look at a few modifications.

It will suffice to consider a partition into three events (such as E_1, E_2 and E_3 of Chapter 3, 3.9.2). We shall call them A, B and C ($A + B + C = 1$) and represent them as points, $A = (1, 0, 0)$, $B = (0, 1, 0)$ and $C = (0, 0, 1)$, in an orthogonal Cartesian system. For the time being, we shall distinguish the probabilities, $p = \mathbf{P}(A)$, $q = \mathbf{P}(B)$ and $r = \mathbf{P}(C)$, attributed to them, from the values x, y and z chosen in accordance with the second criterion (we know they must coincide but we want to investigate what happens if we choose them to be different, either through whim, oversight or ignorance).

---

6 Among other decision criteria that are employed (inspired by points of views which differ from ours) is one which is called the 'minimax' criterion: it consists in taking that decision that minimizes the maximum possible loss. Observe that, in the above situation, this criterion would have us always choose $x = \frac{1}{2}$ (then, in fact, the loss would $= \frac{1}{4}$, with certainty, whereas every other choice would give a smaller loss in one of the two cases, although greater in the other). Since it is incoherent to attribute probability $\frac{1}{2}$ to all events, such a criterion is absurd (in this kind of application; not so, however, in the theory of games – see Chapter 12, Section 12.7.4 – where it provides a solution in situations of a different kind, nor even in this situation under an hypothesis of an extremely convex utility function where it would no longer lead to the choice of $p = \frac{1}{2}$).

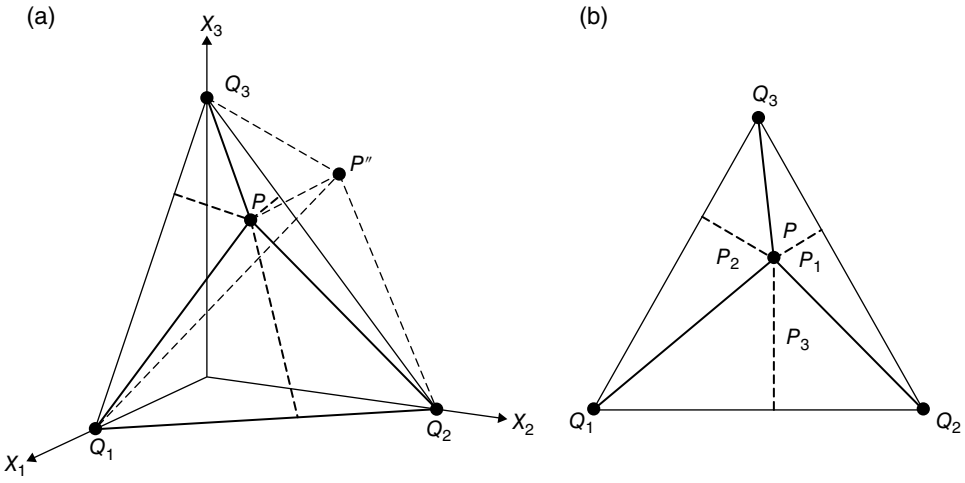


Figure 5.3 The triangles of points such that $x+y+z=1$ (x, y, z non-negative) seen in (a) space, and (b) in the plane. It is clear from geometrical considerations that the choice of a loss rule corresponding to the point (x, y, z) is inadmissible (in the case of three incompatible events) if it is not within the given triangle. Moreover, if one attributes the probabilities (p, q, r) to the three events, it pays then to choose $x=p$, $y=q$, and $z=r$. In other words, the method rewards truthfulness in expressing one's own evaluations.

We shall denote by P the prevision-point $P=(p, q, r)$; the decision-point will be denoted by P'', $P''=(x, y, z)$ (Figure 5.3).

The total loss will then be

$$L=\left(A-x\right)^2+\left(B-y\right)^2+\left(C-z\right)^2 \tag{5.9}$$

and

$$\mathbf{P}\left(L\right)=\left[p\tilde{p}+q\tilde{q}+r\tilde{r}\right]+\left[\left(p-x\right)^2+\left(q-y\right)^2+\left(r-z\right)^2\right]; \tag{5.10}$$

in other words,

$$\mathbf{P}\left(L\right)=\left(\text{first term involving only the prevision}-\text{point } P\right)+\left(P''-P\right)^2,$$

the latter being the square of the distance between P'' and P. Hence, in order to avoid an extra loss, whose prevision is equal to the square of the distance between P'' and P, the point P'' must be made to coincide with P.

The argument given previously (Chapter 3, 3.9.2) was saying the same thing, but without reference to a preselected prevision $\mathbf{P}$. Given a point $P''=(x, y, z)$, outside the plane of A, B and C (i.e. with $x+y+z\neq 1$), its orthogonal projection P' onto this plane has distance less than P'' from A, B and C; if P' falls outside the triangle ABC, the above-mentioned distances decrease if one moves from P' to the nearest point P on the boundary. This shows that only the points of the triangles are admissible (in the sense of Pareto optimality; there are no other points giving better results in all cases). The present argument is less fundamental, but more conclusive, because – assuming the notion of probability to be known in some way (e.g. on the basis of the first criterion) – it shows how and why the evaluations x, y, z of the second criterion must be chosen to coincide with the probabilities p, q, r of A, B and C.

5.4.4. We have here dealt with the most formally immediate case, that of applying to the different events (A, B, C) one and the same scheme with the same maximum loss, namely unity. We know, however (see Chapter 3, 3.3.6), that, so far as the evaluation of probabilities is concerned, and this is what interests us, no modifications would be required were we to use different coefficients: for instance, if one were to take

$$L = a^2(A-x)^2 + b^2(B-y)^2 + c^2(C-z)^2$$

with arbitrary a, b, c. Geometrically, the three orthogonal unit vectors, $A-O$, $B-O$, $C-O$, must now be taken to have lengths a, b and c. This implies – and it is this aspect which may be of interest to us – that the loss, which always equals the square of the distance, is given by $(A-B)^2 = a^2 + b^2$, if in prevision all the probability is concentrated on A, and B actually occurs (and conversely): similarly for $(A-C)^2 = a^2 + c^2$ and $(B-C)^2 = b^2 + c^2$. In the plane of A, B and C, the triangle ABC can be any acute-angled triangle (in the limit, if one of the coefficients is zero, it can be right-angled): in fact, $a^2 = (B-A)\times(C-A) = \overline{AB}\,.\,\overline{AC}.\cos\widehat{BAC}$, $\cos\widehat{BAC} > 0$, and so on. In any case, the scheme would work in the same way even if ABC were taken to be any triangle whatsoever, although if it were obtuse-angled, we could not obtain it as we just did in orthogonal coordinates (merely by changing the three scales). This is obvious by virtue of the affine properties, a point we have made repeatedly. In the general case, the only condition imposed on the three losses $\overline{AB^2}, \overline{AC^2}, \overline{BC^2}$ is the triangle inequality for $\overline{AB}, \overline{AC}, \overline{BC}$.

5.4.5. Why are we bothering about the possibility of modifying the shape of the triangle: that is the ratios of the losses in the different cases? After all, this is irrelevant from the point of view of evaluating probabilities. Despite this, it may sometimes be appropriate to draw a distinction between the more serious 'mistakes', and the less 'serious' (the former to be punished by greater losses), in those cases in which the losses could also serve as a useful means of comparison when considering how things turn out for different individuals (as we shall see shortly).

A good example, and one to which we shall subsequently return, is that of a football match (or some similar game) in which the following three results are possible: A = victory, B = draw, C = defeat. In the most usual case (triangle ABC equilateral), one considers it 'equally bad' if either a draw or defeat results when one has attributed 100% probability to victory. If, on the other hand, the distance between victory and defeat is considered greater than the distance between each of these and a draw, we could take an isosceles triangle with the angle B greater than 60°; if we take this angle <90°, we have a combination of three losses for the three results, and the loss for victory–defeat will be less than twice the loss for draw–defeat (or for draw–victory). For a right angle, this ratio will be exactly double (the ratio of the sides = $\sqrt{2}$) and the scheme will only be applicable to the events victory and defeat (a draw is only taken into account as complementary to the other two). For angles between 90 and 180°, the interpretation as combinations of losses for the three events no longer holds; for the case of a draw, the loss would have to be *negative* in order for things to proceed smoothly! The 180° case means that we are effectively considering prevision in terms of 'points' (0 for a defeat, 1 for a draw, 2 for a victory), in the sense that previsions like (0,1,0) and ($\frac{1}{2}$, 0, $\frac{1}{2}$) – that is of being certain of a draw, or of equal probabilities for victory and defeat, excluding a draw – are considered identical.

5.5 Applications of the Loss Approach

5.5.1. The employment of this method (or something similar) by various people for evaluating probabilities should be given great emphasis and, for many, many reasons, deserves wide publicity.

Sometimes, one is interested in knowing the opinion of a given individual, or of various individuals, concerning the probabilities of certain events under consideration. Sometimes, in order to make some kind of psychological analysis, one is interested in knowing how the various individuals react to information, or other new factors. In certain other cases, it might be interesting to be able to judge, in a more precise fashion, the extent of the 'partial knowledge' of individuals under examination: for instance, one might discourage them from 'guessing'.[7] And so on.

In all these cases, one should take into account the no less important value of repeated experiences of this kind. They greatly aid one in acquiring the 'feeling for numerical values' with which one expresses 'degrees of belief', and hence they contribute to building up a keen and accurate understanding of the problems of prevision, and of the spirit – not cut-and-dried – in which probability theory must approach them.

5.5.2. With this aim in mind, we must now supply all the details of the method. It must be understood that it is preferable to express one's own evaluations sincerely and accurately, and that otherwise one suffers a loss, equal, *in prevision* (in one's own evaluation), to the square of the distance between one's own true evaluation and the one expressed. In addition, there is a definite advantage in obeying the conditions of coherence (in our example; $x, y, z \geqslant 0$, $x + y + z = 1$): to do otherwise is to arrange to suffer one part of the loss *with certainty*. If, instead, one wishes to check – in a decision-theoretic sense – the ability of a given individual to do the right thing without having a systematic knowledge of the situation and of the theory, the characteristic features of the method should not be revealed (except for mentioning what losses are). This is a different problem, however; a far cry from those for which we have introduced the method under present discussion (and it seems unlikely, anyway, that anyone could come to sensible decisions without knowing and applying – with great care! – the theory of probability).[8]

Let us now proceed to some concrete examples of various types of applications.

5.5.3. *The opinions of experts*. It often happens that one turns to the experts for information. This is, in actual fact, nothing other than an evaluation of probability. One is

---

7 By 'guessing', we mean 'guessing at random'. This should not be confused with the usage conveyed in Pólya's 'Let us teach guessing', where it means to make useful conjectures (first guess, then prove!).

8 Experiments of this kind, which are made in order to check the extent to which actual behaviour conforms to the norms derived from the theory of probability, are often considered as 'proving' or 'disproving' the validity of probability theory (or of the related theory of decision making under conditions of uncertainty). This would be so if such theories were to be regarded as empirical–psychological theories of actual behaviour, but, in fact, it is completely at odds with what we are considering here: a *normative* theory for *coherent* behaviour.

Many criticisms derive from this confusion (or from the refusal to accept that a subjectivistic theory can distinguish incoherent and coherent behaviour, rather than just being an acritical, empirical observation of actual behaviour as it happens to be). This kind of empirical evidence is also of interest from our standpoint, but in the same way as a mathematician might find the mistakes of laymen, students, or even other mathematicians, interesting. He does not modify mathematics by incorporating these 'mistakes', as though, simply because someone has enunciated them, they 'should' be included by virtue of their being part of some psychological truth, or of the indiscriminate collection of mathematical statements made in the course of history.

not always in a position to weigh-up for oneself all the probabilities relevant to a given situation; this then is the time to behave like the Prince, who, according to Machiavelli, 'sometimes understands things by himself, sometimes through the understanding of others: while the former is excellent, the latter is also very good'.

An example, one of thousands, is given by the case of a geologist who is asked to give an opinion as to whether it is worth drilling a hole at a particular site during an oil search. This is a useful example to consider, since it has, to some extent, been treated by Grayson,[9] and so the interested reader can delve deeper into those aspects which we shall not discuss. The geologist himself does not have any say in the final decision of whether or not to drill: this decision must be taken (by the 'decision maker') after consideration of all the various pieces of information, of which that of the geologist is just one. He, for his part, cannot state categorically that oil is present or not present (thus making a prediction rather than a prevision), nor can he sin in the opposite direction and merely list the information about the geology of the area (reliable, but analytical), leaving to others the task of synthesis and drawing some conclusions. The synthesizing and the conclusions about the probable outcome of the drilling – given from a geological standpoint – are precisely what his expertise is called upon to provide.

In actual fact, the geologist's report does provide this answer, but usually couched in extremely vague adjectives or phrases (such as: fairly good prospects, or good, favourable, uncertain, promising, etc.; sometimes preceded by little words like 'very', 'not very', 'quite', 'rather', and followed cautiously by 'unless anything unexpected happens', 'perhaps', 'it's difficult to say', 'in my humble opinion',..., 'God only knows'). The only solution worthy of serious consideration is to have the geologist express the probabilities numerically, and some companies actually do this. The objection could be raised (and often is) that the knowledge of the geologist is too vague to be represented numerically. It would certainly be unwise and overzealous to assert that the probability of striking oil at a given site is 0.1307594, but to state that the probability is 0.131, or 0.13, or even simply 10–15%, is always preferable to a string of adjectives whose vagueness depends upon the nature of the opinion itself, on the inadequacy of language, and, perhaps, on a desire to state the conclusions in the least compromising way – that is essentially ambiguous, but not appearing to be.[10]

5.5.4. There remains the problem, however: *how can we interest the expert* – in our case the geologist – *in giving an honest answer; in expressing accurately his deep-felt belief*? This problem was examined by Grayson in the light of the 'first criterion', without any satisfactory solution being obtained. The method we suggest here – that of the 'second criterion' – would seem to give a perfectly satisfactory solution, and is precisely what Grayson requires; 'a system to discourage falsification'. For the practical application at present under consideration, it would be sufficient to agree that some part of the agreed fee (neither insignificant, nor excessive; say, 5–10%) be held back until the eventual outcome was known, and then the loss deducted (up to a maximum of the amount held back) before payment. In certain cases, however, like those of experts who are

9 C.J. Grayson, *Decisions Under Uncertainty*: *Drilling Decisions by Oil and Gas Operators*, Harvard Business School (1960).

10 Someone made the acute observation that often the ability to make accurate predictions consists in expressing them in a sufficiently imprecise fashion (this principle is mentioned on p. 213 of Good's anthology – see footnote 12 – and also in a review article of mine; see *Civilta delle macchine*, No. 1 (1963), 71–72). On the other hand, the limit-case of Sibylline predictions ('*Ibis, redibis*,...') is well known.

consulted regularly, or who hold positions within the firm, one might also add up the losses – expressed as 'scores' – in order to make global comparisons (of the 'goodness' of the previsions of two individuals based on comparisons of the cases examined by both of them). These comparisons could be made separately according to 'type' of problem, time period and so on, and could be taken into account when considering the merits of someone in connection with appointments, promotion and so on.

The following discussion is useful both in real-life and as an example.

5.5.5. *Forecasting sports results.* We consider sports results, football in particular, because they give plenty of scope for experiments of this kind: they can be observed regularly (e.g. every weekend) and sufficiently often; the outcomes are clear-cut (in football, the home team either wins, loses or draws), officially ratified, and the situation is well known to most people. In addition, there is considerable background information and comment in the newspapers. However, leaving aside the convenience (for the reasons given above) of sports results, we could consider forecasting in any area (e.g. politics, economics, meteorology, everyday affairs, culture, judicial or sanitary matters, personal or business affairs, etc.).

There are, as is well known, various organized pools for betting on football and horse racing. These, however, are motivated by the concept of 'prediction', in that they reward those who *guess* all (or almost all) of the *results*. Moreover, the sensitivity of the system is completely distorted by the practice of sharing out the available prize money among the winners. Indeed, the net result of all this is as follows: those who write down ridiculous forecasts, that by chance turn out to be correct, receive fantastic prizes; whereas those who write down forecasts which could reasonably be thought probable receive, if they win, only very small amounts, since the prize, in this case, will presumably have to be shared with many others.[11] Consequently, the 'most reasonable' way to gamble would not be to bet on the result for which the probability of occurrence is highest but, instead, to consider the probability multiplied by the prevision of the reciprocal of the number of people betting on it, and to bet on the result for which this is highest.

The betting approach that we discussed previously, illustrating its merits and demerits, is in line with the notion of prevision (as opposed to prediction). The scheme we are now going to present is intended to build on the merits and eliminate the demerits. It should, therefore, permit us to achieve those goals that we have already mentioned: to develop a feeling for what a prevision (not a prediction) is, and a feeling for the numerical scale on which it is to be expressed; to teach one how to take into account the relevant circumstances, bearing in mind one's own level of competence. Moreover, all this is achieved within the agreeable format of a competition, there being the additional opportunity to reflect and to compare, after the event, one's own previsions with those of others, and with the results themselves. It will be necessary to consider rather carefully the latter point; that is '*being wise after the event*'. We shall do so in Sections 5.9 and 5.10 of this chapter, and will come back to it on several occasions later in the book.

---

11 This brings to mind a rather tragic story: a man died, overwhelmed with joy, on learning that he had guessed correctly all the 13 football results on the Italian football pools. In fact, he was lucky, because otherwise he would have died of disappointment the next day on learning that his winnings were so small (about 3000 Lire), owing to the predictability of the results, which were therefore foreseen by many others besides himself.

5.5.6. One could organize a competition more or less along the following lines (this has already been tried, although on a small scale).[12] The participants have to hand in, each week, previsions for the forthcoming matches, giving, for each match, the probabilities (expressed in percentages) of the three possible outcomes (in the order: win, draw, defeat); writing, for instance, 50–30–20, 82–13–05, 32–36–32 and so on. Given the results, one can evaluate, game by game, the losses and the total losses for the day (and, possibly, a prize for the day), as well as the cumulative sum needed for the final classification. This final classification must be seen as the primary objective. If there are prizes, the largest should be reserved for the final placements, and, in order to conform to the spirit of the competition, the prizes must complement losses; that is they should depend on them in a *linearly* decreasing fashion.[13]

The lessons of experience tell us much about the necessity of avoiding the mentality of prediction when making previsions. It is true that total success – that is no penalty – is achieved if and only if the whole probability, 100%, is attributed to the case which actually occurs. For this reason, many find it tempting, especially at first, to attempt to get the result spot-on, with evaluations which ignore the possibility of uncertainty (i.e. 100–00–00, 00–100–00 or 00–00–100, which are equivalent, in the notation of the football pools, to the 'predictions' '1', 'X', '2'). However, these participants come to realize very quickly that they have fallen behind – this happens on the individual days, but shows up most in the final classification – relative to those who distribute probability in a sensible way: they soon modify their approach.

We shall come back to this example later.

5.5.7. *Replies to multiple-choice questions.* One is often required in a 'quiz', or even in an examination (especially in America), to choose from among a few given answers the one which one believes to be correct. The exact details may differ somewhat: one may either have to tick one and only one answer; or be allowed to choose none; or to choose a subset within which the correct answer is thought to lie (and, in this case, there are two variants, according to whether one indicates an order of preference or not). In any case, there must be an agreed method of scoring according to the way in which the answers given compare with the correct answers. A problem arises from the necessity of discouraging people from 'guessing'; this is often dealt with by estimating statistically what the effect of the assumed presence of 'guessing' would be, in a mass of people.

---

12 It was tried twice, in 1960–1961 and 1961–1962, in the Economics Faculty of Rome University. There were about 30 participants (students and a few teachers) on each occasion, and the study centred on the nine football matches played every week in the first division of the Italian league. Some discussion of this can be found in B. de Finetti, 'Does it make sense to speak of "good probability appraisers"?' in the volume entitled *The Scientist Speculates: An Anthology of Partly-baked Ideas* (edited by I.J. Good) Heinemann, London (1962). The experiment was repeated again in Rome (Faculty of Science) from 1966 on, and experiments of this kind have recently been made in the United States.

13 If, for example, no prize is to be awarded to those who come last (by whatever ruling is proposed), not only do the tail-enders have no motivation to exercise care in their evaluations but, on the contrary, they have a vested interest in trying outlandish evaluations, which they presume to be different from those of individuals in a better position. This is their only hope of overtaking them and getting a prize. If the first prize were extremely large, the temptation to behave in this way would be greatest for those in second place on the next to last day. In any case, such a distortion of interest occurs whenever *linearity* is abandoned.

This latter problem is completely resolved if one applies the method under consideration.[14] Observe that, in this context, there is no question of events which could be considered 'uncertain' in some 'objective' sense. For example: it is clear that if we ask which of A = Antonio, B = Brutus, C = Caesar said the famous line 'Alea jacta est', we are not asking for any sort of testimony or opinion concerning the fact that some great man uttered the phrase in the course of his life; we simply wish to check whether the examinee knows that the phrase relates to Caesar and the crossing of the Rubicon. In the same way, if we ask whether $\log x + \log y$ equals

$$A = \left(\log\left(x + y\right)\right), \quad B = \left(\log\ xy\right) \quad \text{or} \quad C = \left(\log\left(e^x + e^y\right)\right);$$

or whether $\sqrt{26}$ is A = rational, B = algebraic or C = transcendental; or whether, at the battle of Waterloo, Napoleon A = won, B = lost or C = drew; or whether the city of Bahia is in A = Argentina, B = Brazil or C = Chile; and so on; in all these cases, the probability, the doubt to be measured, comes solely from the ignorance, uncertain knowledge or bad memory of the person questioned.

In every other respect, on the other hand, the situation is identical to that of the football pools: for the person who judges, for the person concerned with his state of doubt, there is no difference. It is sufficient to realize that a person could forecast the football results on a Sunday evening, when the facts are part of the past and are known to everybody (provided they are not known to him), or even a year later, provided that he then recollects them with something less than certainty.

5.5.8. The adoption of the proposed system in the case of multiple-choice questions would turn out to be instructive, in addition to the reasons that hold generally (i.e. learning how to express one's own opinion by translating it into numerical values), for the 'lesson' which would show how it is also *advantageous* (where sensible rather than stupid rules are in use) to strive for the greatest honesty and accuracy in expressing one's own doubts or lacunae. Conversely, stupid rules (like stupid laws) encourage dishonesty and reticence, encourage that complex of underhand and stupid actions which are euphemistically described by the phrase 'trying to be clever'; in our case, they encourage 'guessing'.

For the examiners too, it would be extremely useful to have precise information about those who 'know' (e.g. those who write down Antonio 00%, Brutus 00%, Caesar 100%), with the suspicion of 'guessing' now removed, and even more to be able to make a detailed analysis, on the basis of precise and meaningful data, of the frequency, intensity and nature of the doubts (possibly with a view to investigating their origin and suggesting ways of dealing with inadequacies in the teaching). In addition, they would be able to examine the degree of accuracy with which the evaluations are made (e.g. not simply using 50%–50% if there is uncertainty between two alternatives). In the case under consideration, there could, of course, be any number of alternatives whatsoever; in the examples above, we considered three for convenience, and in order to be able to retain the analogy with football results, and the possibility of imagining the situation as always representable in terms of Figure 5.3.

14 The betting approach, on the other hand, could not be used. Anyone in a state of some doubt would certainly lose against an opponent (e.g. the examiner) who knows the right answer.

5.5.9. *Applications in economics.* In the field of economics, the importance of probability is, in certain respects, greater than in any other field. Not only is uncertainty a dominant feature but the course of events is itself largely dependent on people's behaviour, which is itself determined, in a more or less unconscious and confused fashion, by evaluations and arguments of a probabilistic nature. It is, therefore, probability theory, in the broadest and most natural sense, that best aids understanding in this area (and not those fragments of the theory which never progress beyond the drawing of 'equally likely' balls from an urn, or 'stable' frequencies).

This point of view was presented in a clear and authoritative manner by T. Haavelmo in a celebrated critical speech delivered as president of the Econometric Society,[15] where he stated that previsions and evaluations of subjective probabilities '*are realities in the minds of people*' and that it was to be hoped that '*ways and means can and will he found to obtain actual measurements of such data*'

Another point, of particular importance for applications in operational research, is the possibility of making use of those evaluations of probability which represent a decision-maker's own opinions. For example, only the decision maker himself can say what probabilities he attributes to the different reactions of his most direct competitors to possible decisions of his. How, though, are we to interrogate him? Indirect approaches are necessary; questions about his preferences under some hypothetical sets of conditions should be posed in such a way as to provide, in turn, both a complete picture and a check of consistency. These are, however, expedients to make up for the lack of training in expressing oneself in terms of probabilities; the difficulty would not exist if such training became general practice.

Finally (in order not to dwell on too many other aspects[16]), there are important applications to the more theoretical field of econometric models. As E. Malinvaud says, in his treatise on statistical methods in econometrics,[17] the justification of the introduction of random models into econometrics rests, in his view, on an appeal to subjective probabilities, so that 'l'établissement d'une statistique subjectiviste qui reposerait sur le principe de Bayes' would be desirable (even though, in his opinion, research in this direction is, as yet, not sufficiently advanced to make a systematic application possible: on the other hand, there are those, for example A. Zellner,[18] who are attempting to do this).

5.6 Subsidiary Criteria for Evaluating Probabilities

Having analysed the meaning and the method of evaluating probabilities that a person might be led or compelled to make in order to sort out his ideas about what might occur, and to choose wisely any decision that has to be made, we are now in a position,

---

15 Trygve Haavelmo, *The rôle of the econometrician in the advancement of economic theory*, Presidential Address, Meeting of the Econometric Society, Philadelphia, 29 Dec. 1957; see *Econometrica*, **26** (1958), 351–357.

16 I have recently provided a wide ranging discussion of these topics (with a fairly mathematical treatment) in 'L'incertezza nell'economia,' part I of: B. de Finetti and F. Emanuelli, *Economia delle assicurazioni*, Vol. XVI of *Trattato italiano di economia* (Edited by C. Arena and G. Del Vecchio), Utet, Torino (1967).

17 Edmond Malinvaud, *Méthodes statistiques dans l'économétrie*, Dunod, Paris (1964).

18 Arnold Zellner, *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons (1970). It should be noted, however, that, although the treatment is Bayesian, the interpretation is not subjectivistic. The choice of the initial distribution does not derive from a case-by-case consideration of the factual circumstances, but from adopting once and for all a mathematically convenient form for each type of problem.

and are in fact obliged, to return to the essence of the problem of evaluation. We wish to discover whether the task of translating more or less vague impressions and opinions into numerical form could be facilitated by using some suitable subsidiary criterion. Fortunately, this turns out to be the case.

This happy circumstance derives, in general, from the observation that in many cases in the calculus of probability, under restrictions that are often very natural, certain probabilities, which are calculated on the basis of certain others, vary very little as one's evaluations of these other probabilities are varied. Consequently, even if the latter seem, to a given individual, rather vague, the former may very well appear to him capable of being evaluated with sufficient precision and confidence. As a brief aside on the question of interpersonal comparisons, we note that this explains why individuals often make practically identical judgments of prevision, even though they start off with very different opinions.

These general considerations will become clearer as we proceed further. For the time being, we restrict ourselves to illustrating the two subsidiary criteria which are of the greatest and most immediate interest: the first one we shall deal with in a reasonably detailed manner; the second, which, from a logical point of view, is based on material we shall meet much later on, is dealt with in a necessarily superficial way.

5.7 Partitions into Equally Probable Events

5.7.1. Every quantitative measurement is made both easier and more precise when it is possible to reduce it to a qualitative comparison. For example; it is much easier to say that A.N. Other has eaten $\frac{2}{9}$ (i.e. about $22 \cdot 2$ %) of a cake knowing that it was divided into 18 pieces, which could be taken as equal, and that he has eaten four of them, than to directly estimate that his portion was $22 \cdot 2$ % of the whole, undivided cake. In precisely the same way, it is obvious that if I judge n events of a partition to be equally likely, I cannot avoid attributing probability $p = 1/n$ to each of them (because the sum of the n terms, each equal to p, must be 1). Judgments of this kind arise rather frequently: it is sufficient that, given the present state of information, one finds oneself in a situation of *symmetry*. This will often, although not necessarily always, reduce to a state of *symmetry* regarding certain physical, or at any rate external, circumstances, which we regard as essential and relevant elements of our state of information.

When tossing a coin, we usually attribute the same probability $\frac{1}{2}$ to both faces and, similarly, probability $\frac{1}{6}$ to each of the six faces of a die. If we have n balls in an urn, we again, in general, attribute the same probability $1/n$ to any particular one of them being drawn: in this case, if we also know that m of the balls are white, we have no choice but to attribute probability m/n to the drawing of a white ball. This judgment of equiprobability (relative to a single toss, throw or drawing – this is not the place to consider more complicated cases) reflects a symmetric situation which is often made objectively precise by stating that the balls must be identical, the coin and the die perfect (physically symmetric) and so on. However, the criterion remains essentially subjective, because the choice, of a more or less arbitrary character, of those more or less objective requirements which are to be included in this concept of 'identical', reflects the subjective distinction drawn by each individual of what is, and what is not, a circumstance that influences his opinion. It was necessary to point this out, in order to avoid giving the impression that in problems of this kind we are dealing with a different kind of probability; objective rather

than subjective. It is true, however, that in this context opinions generally do coincide (although the agreement is less strong and unconditional than one would tend to think). Independently of all this, we can always talk about the case of equiprobability, provided we state (or take it as implicit) that this simply means that You (or the individual concerned) attribute the same values to the probabilities in question.

5.7.2. Returning to our examples, we observe that by means of these kinds of set-ups – it might be sufficient just to consider drawings from an urn – we can easily obtain a representation of events of any given probability (to be more precise, any rational m/n). For example, if one wants to get an idea of the magnitude of a probability expressed to two or three decimal places, for example expressed in percentage terms like 13% or $13 \cdot 2\%$, it is sufficient to think of an urn with 100 balls, 13 of which are white (or 1000 balls, 130 or 132 of which are white). One can avoid talking about colours, and changing the percentage of white balls, by simply thinking of the balls as numbered consecutively (from 1 to 100 or 1 to 1000): this enables one to say – albeit in less suggestive language – that 13% is the probability of drawing a number not exceeding 13 (out of 100; or 130 out of 1000) and so on.

Using the 'representations' of this 'scale' one can – if the method seems easier – reduce the evaluation of any probability to comparison with cases of this kind, and forget all about both the betting approach and that in terms of losses. In order to translate into figures the probability – according to You – of striking oil by drilling at a given spot, it is sufficient that You decide how many balls, out of 1000, should be white, in order to obtain the same probability of drawing a white ball; if You think the number should be 131, this implies that You think the probability of striking oil is $13 \cdot 1\%$.

It is convenient to express all this formally:

Theorem. *If the n events of a partition are considered as equally probable, the probability of each of them is 1/n, and the probability of an event which is the sum of m of them is m/n.*

The classical statement is that, under these conditions, *the probability is given by the ratio of the* '*number of favourable cases*'(m) *to the* '*total number of possible cases*' (n).

5.7.3. *Criterion of comparison* (or 'third criterion' – following the two in Chapter 3, Section 3.3). Having at one's disposal a model of a partition into n events, which are judged equally probable (e.g. an urn), *the probability of any event E can be evaluated, by comparison with events composed of sums of events of the partition, with an error of less than* $1/n$. In fact, if E_m and E_{m+1} are sums of m and $m+1$ events, each of probability $1/n$, and if one judges $\mathbf{P}(E_m) \leqslant \mathbf{P}(E) \leqslant \mathbf{P}(E_{m+1})$, then $m/n \leqslant \mathbf{P}(E) \leqslant (m+1)/n$. In order to make the comparison operational, it is sufficient to express it by saying that You would rather receive one lira if E occurs than one lira if E_m occurs, but vice versa if the comparison is made with E_{m+1}. In this way, its subjective nature is clear; it remains somewhat in the shade when we speak of 'comparison' in the abstract, with no precise meaning.

There are many points, both historical and critical, that one could raise at this juncture, but they would require overlong, and in part untimely, digressions: they will be considered instead at the end of the Appendix.

Let us just say something, however, in order to make the above a little more precise, at least in its essential features. Evaluations made on the grounds of symmetry are generally accepted as a basis for problems concerning games, drawings from an urn, lotteries, dice and so on, and one often regards as 'equally probable cases' certain outcomes which are

'combined' (like the 6^{10} possible sequences obtained by tossing a die 10 times, or the 90 !/85! possible sets of five numbers on the Lottery, or the 90! permutations in a drawing of all the 90 numbers at Bingo etc.), rather than elementary (like the score obtained at the next throw of a given die, or the number 'drawn first' at a given Lottery wheel next Saturday). Recall the remarks made in Chapter 4, 4.10.3, which are relevant to this procedure.

5.7.4. We note, however, that it is not just in examples of games of chance that considerations of symmetry can act as a guide but, in fact, in any practical problem whatsoever. For example: if we consider the maximum annual temperature (at a given location) in three consecutive years, then it can either:

- increase (type 1–2–3, where 1, 2 and 3 schematically denote the three temperatures in increasing order),
- decrease (3–2–1),
- be maximal in the middle year (types 1–3–2 and 2–3–1), or be minimal in the middle year (types 2–1–3 and 3–1–2).

Now, whatever one's evaluations of the probabilities of more or less high summer temperatures might be, under certain conditions it may very well be natural for us to attribute the same probability $(\frac{1}{6})$ to each of the possible cases.

Example A. *Increases and decreases in agricultural production*. This is a (true!) example of a fallacious analysis, based on the observation that, by comparing agricultural production in successive years, the numbers of *inversions of trend* (i.e. the number of times in which an increase was followed by a decrease, or vice versa) was about twice the number of *permanences* (i.e. repetitions of an increase or of a decrease). An agricultural expert argued that rich and poor crops alternate, and it required a statistician to point out the mistake (the numbers are in agreement with what we have just seen above).

Example B. *Breaking an existing record*. In connection with temperatures, agricultural production, or even the results in an annual competition, for example the winning throw in the national discus championships (assuming the given hypotheses continue to hold: i.e. there exists no reason to expect an improvement due to better training, more participants etc.), one can pose the following sorts of problems: what is the probability that in the nth year (of the competition, of keeping temperature records etc.) a new record is set up? (Ans. $1/n$); that the record (set in the first year) be broken for the first time? (Ans. $1/n(n-1)$); that the previous record had stood for h years ($h = n-1$, $n-2$,..., 3, 2, 1)? (Ans. $1/(n-1)$ for any h); what is the prevision of the number of times the record was broken in the first n years? (Ans. $\Sigma(1/h)(1 \leqslant h \leqslant n)''$ $\log n$); and what is the prevision of the number of years that the record lasts until the next improvement? (Ans. $+\infty$). As an exercise, verify these answers and pose yourself some further problems (these are easy to find, although not always easy to solve).

5.8 The Prevision of a Frequency

5.8.1. When considering events $E_1, E_2, \ldots, E_n$, it may happen that we know with certainty what the number of successes $Y = E_1 + E_2 + \ldots + E_n$ (or, equivalently, the frequency Y/n) must be: $Y = y$, say; that is $Y/n = y/n$. Clearly (see Chapter 3, 3.10.3), the sum of the

$p_i = \mathbf{P}(E_i)$ must be equal to y (i.e. their arithmetic mean must be equal to y/n); in particular, if the E_i are judged to be equally probable, $p_i = p$, then we must have $p = y/n$ (the probability equal to the known frequency: for $y = 1$ we have the case of a partition, as considered previously). However, even if the frequency is not known with certainty, the relation still holds if we substitute the prevision of the frequency: *the sum of the probabilities must equal the prevision of the number of successes*. In other words, dividing by n, we have

Theorem. *The arithmetic mean of the probabilities must equal the prevision of the frequency:*

$$\left(p_1 + p_2 + \ldots + p_n\right)/n = \mathbf{P}\left(Y/n\right) = \mathbf{P}\left(Y\right)/n. \tag{5.11}$$

In particular, *if the E_i are judged equally probable, $p_i = p$, we have $p = \mathbf{P}(Y/n) = \mathbf{P}(Y)/n$: the probability* (*common to all the events*) *is equal to the prevision of the frequency*.

5.8.2. In order that correct use be made of this theorem, we must make very clear that it is essentially trivial: otherwise, we run the risk of goodness knows what being read into it. Observe first of all that the E_i can be any events whatsoever, however diverse, so long as the number of successes is given by addition: for example success in an examination, a victory for one's favourite football team, finding a traffic-light green, throwing a double six at dice, and anything else, however dissimilar. The 'theorem' is an identity: it imposes no restrictions, apart from informing us that the same thing, written in two ways, remains one and the same thing (rather like the sum of a double-entry table, which can be taken either over rows or over columns).

Well then: *it is in this very thing – and in nothing else – that the value of any theorem in the calculus of probability lies, and it cannot be otherwise. It is to tell us whether, in making the same evaluation in two different ways, we arrive at different conclusions, and, in this case, to invite us to think again and to rectify the situation by modifying one or the other*.

There is no *unique way* of doing this: we do not begin with one side already fixed and the other to be 'deduced'. Instead, we have on both sides evaluations that should agree, and which must be modified if they do not. How should this be done? Generally speaking, one of the evaluations usually seems to be more immediate, so one is inclined to look for a modification of the other; however, one should be open-minded about it, since appearance might well be only appearance.

5.8.3. Turning now to our particular case, You might find that the probabilities which You have evaluated, when added together give, for instance, a value which is greater than the number of successes, $\mathbf{P}(Y)$, which, in prevision, seems to You reasonable. You must then ask yourself: 'have I given the p_i values which are on average too large, or are the values which I thought of for the number of successes Y (or the frequency Y/n) too low?' It is fairly difficult to answer this if the events are rather disparate, but when they are more alike, and especially if we know the frequency of other similar events, which have already been observed, it often happens that one places greater confidence in prevision of the future frequency (under the assumption that it will remain close to that previously observed).

Why is this so? The answer to this cannot be given at present (see Chapter 11) but, even without going into the whys and wherefores, the idea that there is a degree of stability in the frequency of occurrence of events usually grouped together as 'similar' is one which

seems quite intuitive to most people. At the present time, this phenomenon may even be somewhat exaggerated as a result of overly simple and rigid formulations current among many statisticians. However, it rests on a very real foundation, since this is how things appear, even to the naïve layman (who, for example, is really surprised if in a given period certain phenomena re-occur with an unusual frequency). Let us accept things as they are.

As a particular case, suppose the events under consideration are so similar that one judges them equally probable: it will turn out that their probability p will be evaluated on the basis of a frequency f observed among similar events in the past, and that p will be close to f. Notice that in this case the evaluation is based not only on the prevision of a frequency, *but also requires a judgment of equal probabilities*.[19]

5.8.4. *Some examples.* Statistics show that the percentage (or frequency) of males among live births is always about 51·7% (hence, a few more males than females); that, according to the Italian tabulations for 1950–1953 and 1954–1957, respectively, the percentages of deaths in the first year were 6·75% and 5·49% for males, 5·88% and 4·67% for females; that the overall annual percentage of deaths in Italy in 1960 attributable to cancer was 1·51%, but broken down into age groups it was

Age:	0–5	5–25	25–55	55–75	over 75
%	0·013	0·009	0·078	0·524	1·131

and into regions (not distinguishing age groups) it varied from 0·220% in Liguria, 0·210% in Tuscany, to 0·089 % in Puglia and 0·073% in Basilicata and Calabria. To change the subject completely, statistics also show that the results of championship football matches are distributed (in terms of home fixtures) as 50% wins, 30% draws, 20% defeats.

Thinking of such frequencies as stable, we could adopt them universally as probabilities for any similar events, or future cases; or, at least, we could evaluate the probabilities of individual cases in such a way as to make them compatible, in arithmetic mean, to these frequencies. However

5.8.5. *The need for realism.* Even though we have expressed our previous considerations with a certain amount of caution (which itself might appear overdone and unnecessary to anyone accustomed to a different approach), it is necessary, in fact, to go still further and provide additional warnings in emphasis of that caution. We seek to reduce everything to three questions (and in answering these we shall delve deeper).

5.8.6. *The first question*: are we justified, in real applications, in attributing the same probability to all the events of a given type? This question is equally relevant to both of the subsidiary criteria; that is symmetric partitions and frequencies. However, we must first point out that it is meaningless unless we bear in mind that the probability is not an *external fact* relating to the event, but, instead, relating to your state of information regarding the event, and the previsions which You derive from this state of information. If You know the innate qualities, the past records and the degree of preparedness of

19 This is often overlooked: if, for example, one speaks of 'the probability of a newly born baby being a boy', it is not made explicit that one is dealing with one unspecified event out of infinitely many ill-defined events, each of which is understood to be equally probable.

every student, your evaluation of the probability of passing an examination will vary from student to student. Even with all this background information, however, if You only know the students by sight (i.e. are ignorant of the name of any given student) and are asked name by name to give the probabilities, then your evaluations will all be equal (the same would be true if knowledge by sight or by name were the other way around). In much the same way, your probabilities for the results of different football matches on a given day will be different if You know the merits of the respective teams, and are in a position to express a prevision for each match. However, if You had to fill in a pools coupon knowing what the matches were, but not the order in which they were listed, You could only assign the same probabilities to them all (the averages of those for the individual matches). For example, it might be 40–20–40 if in about half of the matches the away teams are first-rate and favourite to win, or, if You had to fill it in without even knowing what matches were being played that day, You might adopt a standard average probability like 50–30–20. Even in the near legendary case of drawings from an urn, for instance drawing one from among 90 identical balls (numbered from 1 to 90), equality would not necessarily hold if one knew the position of each ball in the urn at the instant before the drawing took place (You might know, or believe, that the person drawing the ball has a habit of drawing from the top, or from the left-hand side, and so on, and taking this into account might lead You to judge the probabilities to be different).

5.8.7. *The second question*: if I wish to make use of a frequency, which one should I base my opinions on? Given an event E in which You are interested, there are usually several classes of events already observed, which are, in different ways, more or less directly similar to yours, but with each class providing a different frequency: the choice is largely arbitrary.

Let us consider, for example, the problem of life insurance for a certain individual (for simplicity, suppose it is a question of a capital sum being provided if he dies within a year). How shall we determine the 'premium'; that is the probability of his death within the year (not taking into account any 'extras' – for expenses, etc.). We could check the statistics of the deaths of individuals in the same country (or region, county, city, district, etc.), of the same age (sex, class, etc.), having the same profession (income, degree, etc.), of similar constitution (height, weight, etc.), same name or initial of surname, or house number, or born in the same month, and so on: or we could group together some number (large or small) of these sort of characteristics, or any others. Each grouping will yield a different frequency, and this forces us to adopt a reasoned evaluation rather than a mechanical one; one which takes into account those classifications which it appears most reasonable to assume related to the phenomenon (for instance, age), and not the others (like the person's name). What is 'reasonable' depends not only on whether and *to what extent* this or that circumstance influences the phenomenon, but also on *how* it has an influence. If, for example, it appears reasonable (on general grounds, and on the basis of corroborative evidence) to think that the death rate increases with age (once we pass childhood), one would be inclined to stick to this when evaluating the probabilities of death in the immediate future, even for those countries for which the most up-to-date statistics would show oscillations from year to year. One would appeal to some sort of *smoothing* procedure, in an attempt to preserve the general outline, which is considered significant, and to eliminate what are thought of as misleading perturbations.

Finally, one is always faced with the aspect we have already spoken of; that of individual differences (which the insurance companies take into account through the results of the medical examination).

This is a general situation and examples are easy to come by. We shall consider just one other, which shows how meaningful variations in frequency, for appropriately chosen subdivisions, can occur, even in those situations where it appears to be more correct to view the probability as invariant with respect to any of the background circumstances. Given that the frequency of males among new-born babies is almost completely invariant over time, races, or countries, there would seem to be no possibility of differentiating probabilities on the basis of frequency statistics selected according to some factor or other. On the contrary, the research of Gini (using Geissler's data on Saxony, 1876–1885) brought to light a differentiation on the basis of families: there were too many families with an excess of either males or females for it to be 'attributed to chance'.[20] Presumably, one could always find some differentiation if one could succeed in finding appropriate factors on which to base a classification. On the other hand, clearly, as a kind of converse, for those for whom every attempt at picking out significant factors is unsuccessful, every combination of cases automatically appears uniform (even if this is not so for those who do succeed in picking out such factors).

5.8.8. *The third question*: are we justified in expecting frequencies to be stable? The remarks concerning the second question have already led us to consider the differences in frequencies when we refer to subgroups (e.g. in questions concerning people, age groups, regional groupings, etc.), not to mention individual differences (as discussed in the first question). The stability of all these frequencies is an hypothesis, incompatible with the variability exhibited by the overall composition in terms of subgroupings (e.g. dividing the population according to age, region, etc.). In actual fact, in practice, we can usually assume that the overall composition changes rather slowly and, therefore, that the incompatibility is not obvious over a short time period: from a logical point of view, however (and in some cases from a practical one too), the objection is completely valid. On the other hand, even if we leave all this out of consideration, there may be – and there usually are – causes of variation resulting from the evolution of the situation itself. For example, if we consider mortality, there has been great progress in sanitation, medicine, general living standards and so on, as a result of which mortality has progressively declined significantly (this can be seen even from the snippets of data we reported above, relating to very close time periods like 1950–1953, 1954–1957). It might, therefore, appear reasonable, in evaluating a future probability, to extrapolate the rate of improvement rather than base oneself on the hypothesis of the preservation of the present level.[21] In any case, the force of the 'stability of frequencies' as a probabilistic or statistical principle is completely illusory, and without solid foundation.

Similar considerations apply, of course, in other fields. We could add the obvious examples of frequencies of car accidents, and similar matters in connection with

---

20 C. Gini, *Il sesso dal punto di vista statistico*, Sandron (1908), Ch. X, 'La variabilità individuate nella tendenza a produrre i due sessi' (pp. 371–393). I do not know whether there has been any more recent research confirming these results: in any case, it is the argument which is of interest here rather than the facts.

21 Questions of this nature have been discussed with particular reference to the actuarial field; see R.D. Clarke, 'The concept of probability', *J. Inst. Actuaries*, 1954.

technical or economic development. In the case of football, the changing character of systems of play, tactics, and many other things, may alter the influence of playing at home or away, and therefore the probabilities of the three results. In addition, even without changes of this kind, frequencies will be altered if the imbalance between 'top' teams and 'bottom' teams is altered.[22]

5.9 Frequency and 'Wisdom after the Event'

5.9.1. Let us repeat an earlier remark, whose function is to prevent a certain confusion; one which we have already warned against, but to which we are particularly vulnerable in the case of previsions of frequencies.

Previsions are not predictions, so there is no point in comparing the previsions with the results in order to discuss whether the former have been 'confirmed' or 'contradicted', as if it made sense, being 'wise after the event', to ask whether they were 'right' or 'wrong'. For frequencies, as for everything else, it is a question of prevision not prediction. It is a question of previsions made in the light of a given state of information; these cannot be judged in the light of one's 'wisdom after the event', when the state of information is a different one (indeed, for the given prevision, the latter information is complete: the uncertainty, the evaluation of which was the subject under discussion no longer exists). Only if one came to realize that there were inadequacies in the analysis and use of the original state of information, which one should have been aware of at that time (like errors in calculation, oversights which one noticed soon after, etc.), would it be permissible to talk of 'mistakes' in making a prevision.

Any reluctance one feels in accepting these obvious explanations is possibly accounted for by their seeming to preclude any possibility of taking past experiences into account when thinking about the future. This is not so, however: the latter is rather different from 'correcting previous evaluations'. One must emphasize that this phrase is wrong, even though it may only be a confused way of expressing an actual need. It is not, however, a harmless inaccuracy: in actual fact, it distorts the basic question, and generates a tangle of confusions and obscurities.

This should be made absolutely clear. If, on the basis of observations, and, in particular, observed frequencies, one formulates new and different previsions for future events, or for events whose outcome is unknown, it *is not* a question of a *correction*. It is simply a question of a new evaluation, *cohering with the previous one*, and making use – by means of Bayes's theorem – of the new results which enrich one's state of information, drawing out of this the evaluations *corresponding to this new state of information*. For the person making them (You, me, some other individual), these evaluations are as correct *now*, as were, and are, the preceding ones, thought of *then*. There is no contradiction in saying that my watch is correct because it now says 10.05 p.m., and that it was also correct four hours ago, although it then said 6.05 p.m.

5.9.2. Discussions and refinements of this kind, which might seem rather pointless when made in the abstract and reduced to mere phrases, are not only of genuine

22 If, for example, one half of the teams were so much stronger than the others that they beat them with certainty, then about half the matches would have the assigned result; if the frequencies 50–30–20 were retained for the other half, we would have, overall, the frequencies 50–15–35 (the averages of 50–00–50 and 50–30–20).

relevance to the conceptual and mathematical construction of the theory of probability, but they also have implications which demand the attention of everyone; even those not interested in topics of this nature.

The meaning which attaches to statements about 'being wise after the event' does not seem to correspond in a unique way to attitudes either for or against the considerations just made. It is often both different and opposite. This happens when the sentence is uttered as a reproach to someone who belatedly admits that 'he was wrong' – as if to tell him '*tu l'as voulu…*' ['it is your own fault…']. It is conceivable that in some situations this reproach is justified: one often makes mistakes through lack of concentration, or because one was unable to resist the temptation, although fully aware of being in the wrong.

However, the reproach is often made when there is no fault – apart from that of failing to be a prophet. Judgment *by results*, the notion that someone's merit should be measured in terms of his successes, is often passed off as 'realism': to dwell upon the *ifs* and *buts* is considered meaningless. Of course, it is meaningless as far as the facts are concerned; no-one doubts that these cannot be reversed or modified by any *ifs* and *buts*. The facts themselves are not open to question, but when we turn to *judgments based on those facts*, *evaluations of personal responsibility*, *appreciation or criticism of someone's actions*, it is a different matter. In these matters, it is by no means true that the facts provide any definite answers; in fact, they provide no answers at all. Their only value might be in helping one towards a better understanding of the range of ifs and buts. It is precisely these which allow one to judge someone's actions in the one way that makes any sense: that is taking into account, moment by moment, *the context*, *the situation and the state of information in which the actions took place*.

It would perhaps be overstating the case to suggest, for these reasons, the removal of any distinction between – let us say – being found guilty of murder and of attempted murder. It could happen that 'missing' killing someone was evidence of a lesser intention of doing so; but if everything hinges on a miraculous piece of surgery, how is the offence in any way less serious, or the culprit more deserving of leniency? Anyway, since legal matters are somewhat of a mystery to me, I do not wish to pursue the question.

Something that can be criticized with more certainty is what seems to me the deplorable habit of picking on someone as a scapegoat when something goes wrong. Apart from being unfair, the practice encourages people to avoid taking on responsibility, so that one gets the worst of all worlds. Those who acted loyally, in a sensible manner, cannot be reproached if, by chance, the outcome was unfavourable; those who blundered (in an honest fashion) are advised to learn from the experience and take more care in the future. In contrast, those who had not done everything possible, in terms of organization, control and efficiency, to reduce the risk of unfavourable outcomes are punished – whether or not anyone was responsible.

To set against such stupidity, there is an alternative practice, which can be taken as an example of the beneficial effect of a mode of thinking based on operational research. It was brought to my attention by Pasquale Saraceno,[23] and is established practice in the industrial group of which he is one of the leading figures. When examining the actions of the various companies, and especially those with unfavourable outcomes, the analysis is based on drawing a distinction between that which could and should have been foreseen, on the

23 *Translators' note.* Italian economist; former head of the I.R.I. (the state controlled Institute for Industrial Reconstruction).

basis of the information at hand, and that which could not possibly have been foreseen. This sort of calm criticism and self-examination is undoubtedly what is required in order to encourage a sense of responsibility in a climate of honesty and mutual confidence.

5.9.3. The remarks above were made in order to underline the importance of breaking away from these destructive hangovers of the confusion between prediction and prevision: this is important from a general – one might even say *moral* – point of view. Let us turn to a technical aspect of the problem, which should help to remove such confusion. I say 'should', because I know only too well that such errors (these, it seems, more than most) are difficult to eliminate; like the Hydra with a thousand heads. Were it not for this, I would have simply said, as it seems to me, that each objection raised is decisive in itself, and should be sufficient.

In order to combat the idea that the influence of the facts, or, to be more precise, of information regarding the facts, on prevision should be interpreted as a mechanism for refutation and correction (and also to point out the inadequacy and awkwardness of language which gives this impression), we observed that the 'new' opinion, far from being new, was already contained in the 'old', which, far from being refuted, was used when we took over as the 'new' the opinion it had already provided as appropriate for such an eventuality (as for any other possible outcome).

Let us note at this point that such an 'opinion implicitly contained' in the initial one, and already provided for such a contingency, is integrated with it to such an extent that it practically lends itself to being used *without even the occurrence of the facts under consideration*.

5.9.4. The '*device of imaginary observations*', put forward, in particular, by Good (1950), is a method of evaluating probabilities, and, as such, deserves mention in the present chapter. It is a device that is particularly useful for evaluating very small probabilities, and which is more accurate in this context than the direct approach. A simple example will suffice to make the notion clear.

A person claims that he is able to guess in which hand you are concealing a certain object; You do not believe him. If You are invited to be more precise and say what probability p You attach to the possibility that he really can do what he says, You reply 'very small', but are not really in a position to sort out the different implications of saying 10^{-2}, or 10^{-10}, or some other value. Then, according to Good, one can do better by reformulating the question in the following way. Imagine that You put him to the test, and that he guesses correctly three times in a row, or ten times, or fifty times, …; after how many consecutive correct guesses would You consider as equally likely (probabilities $\frac{1}{2}$ and $\frac{1}{2}$) the two possibilities that either his claims are justified, or that he has guessed correctly by chance?

It is easily seen that at each trial where he guesses correctly the probability ratio in favour of his claims doubles (likelihood ratio $1{:}\frac{1}{2} = 2{:}1$); after n such trials it is 2^n. If, after n trials, the ratio of the probabilities has become $\frac{1}{2} : \frac{1}{2} = 1$, it must mean that initially it was given by $p{:}\tilde{p} = 2^{-n}$; in other words, we have approximately, $p = (\frac{1}{2})^n = 10^{-n\log 2} \approx 10^{-(0{\cdot}3)n}$. For example: if $n = 10$, we have $p = 10^{-3} = 0.1\%$; if $n = 30$, $p = 10^{-9}$; if $n = 50$, $p = 10^{-15}$.

There is no doubt that, with this interpretation available, a comparison between the meaning of answers such as $p = 10^{-3}$,or $p = 10^{-100}$, is no longer unattainable (although a certain vagueness or unfamiliarity is inherent in questions of this kind, and cannot be removed altogether; any method or device of this kind is intended as an *aid*, not a panacea:

once one gets to a certain stage, there is nothing to do but try to sharpen the feeling for numerical values of probabilities, including the very small ones).

The conclusion regarding the principle of this method seems to derive further support, psychologically speaking, from a consideration of the paradoxical – I would even say grotesque – position that a contrary point of view leads one into. Its formulation would have to be along the following lines (any attempt at spicing it up in order to increase its paradoxical and mind-bending flavour would only spoil it):

'*My initial evaluation was* $p' = (\frac{1}{2})^n$;

'*it was based on consideration of a hypothetical possibility; that of a succession of n experiments, in which the person claiming to be able to guess obtains successes on every trial, and on my reaction to such a hypothetical result, consisting precisely in the fact that my final evaluation would then have been* $p'' = \frac{1}{2}\tilde{p}$;

'*now, the eventuality considered as the hypothesis has actually occurred, and my reaction has been precisely the presupposed one, therefore…*

'*the initial evaluation, which was, and still is, a logical consequence of these assumptions* (*actual or hypothetical*)*… WAS FALSE*!'.

5.10 Some Warnings

5.10.1. It is necessary to point out a number of pitfalls. Although it is premature to talk about the dangers before we understand their causes, some pointers must be given in order to guard against the doubts and distortions that might get mixed up with what we have said concerning the evaluation of probabilities, giving rise to confused and contradictory notions.

The following remarks should, in one sense, be unnecessary. All the dangers have already been mentioned and the details already given at the appropriate time would be sufficient to render these additional comments superfluous, if – and this is the difficulty – they remained firmly implanted in one's mind, together with all their ramifications, and with such clarity that any dangers reappearing, in whatever disguise, could be dealt with just as effectively as when they were first encountered. This not being the case, it is preferable, and perhaps necessary, to repeat ourselves; to go over the details mentioned above, in their different variants and versions, pointing out the many forms the dangers may assume. (There are such a number of them that perhaps some, even important ones, will be overlooked; hopefully, though, the pattern-book of objections and counter-objections will be sufficiently representative for the reader to be able to answer, by analogy, possible objections not covered, by means of suitable counter-objections.)

5.10.2. It might be argued that the kind of problems we have considered in this treatment, and for which we have discussed the appropriate methods of evaluating probabilities, are outside the 'true' ambit of the calculus of probability, or, at most, they constitute a small and specialized part of it.

The arguments put forward will be, by and large, the standard ones; however, if they are given with reference to physics, for instance, they may appear novel, or at least more substantial and difficult to refute.

There are cases where the probabilities in physics are given by combinatorial arguments, in accordance with the 'classical' idea of 'equally likely cases'; that is, they are given by

the Maxwell–Boltzmann, Bose–Einstein, Fermi–Dirac 'statistics' (to use the jargon of physicists): for further details, see Chapter 10, Section 10.3. Who can argue in this case that we are dealing with a probability whose value is objectively determined by '*a priori*' considerations? It is precisely this example (as Feller observed, vol. I, pp. 5 and 21) which shows how fallacious any *a priori* conclusion would be: nobody could have foreseen that the computation of 'equally likely cases' had to be carried out using completely different methods in problems where different 'statistics' apply (and the explanation only came later, through the distinction between particles with integer or semi-integer spin).

5.10.3. Everyone will probably agree, therefore, that it makes no sense to be willing to deduce properties of phenomena, or previsions regarding their outcomes, basing oneself solely on superficial, preconceived ideas. The confirmation of experience is required, and this, certainly, leads on to an objective conclusion. One might well say that, for the physicist, probability coincides with frequency.

And this statement is, in a certain sense, true. However, this form of expression is completely wrong from a conceptual point of view, even if at first sight it presents no difficulties.

Let us swiftly demolish, one by one, the main arguments put forward with the intention of transforming probability from being subjective to being objective, by means of more or less overt confusion or connection of the notion with that of frequency.

5.10.4. Firstly, we present an objection frequently raised against the notion of the *probability of an individual event*: either this event occurs or it does not, and therefore it either has probability *one* or *zero*; it makes no sense to attribute to it an intermediate probability p. I accept this argument completely, in that it refers to an objective probability p: but I observe that the same argument holds even in cases where my opponent forgets that it does – when he says that in n 'individual cases' there is an objective meaning to p because np of them will occur. This is not true: either *zero*, or *one*, or *two*,..., or *all n* of them occurs, and the objective probability (if one prefers to use this term as a useless and misleading synonym for frequency) is one of the $n+1$ values $0, 1/n, 2/n,\ldots, h/n,\ldots, (n-1)/n, 1$, although it is not known which one.

It is only in a subjective sense that it makes sense to speak of p, as the arithmetic mean of these $n+1$ possible values, *taking as weights the subjective probabilities of the single frequencies* (*still* '*individual cases*'!).

5.10.5. It might be objected that in many cases (those to which an opponent would limit himself) the probability is concentrated near a certain frequency p, which could be defined as objective probability. But here, and in every case in which something 'very probable' is said to be 'practically certain' (or even 'certain', for the sake of brevity), and, symmetrically, something 'very improbable' is said to be 'practically impossible' (or even 'impossible'), an *either–or* must be clearly established. In fact, such sentences can either say something obvious, with which one has no choice but to agree, or, alternatively, they can completely falsify the meaning of things. The field of probability and statistics is then transformed into a Tower of Babel, in which only the most naïve amateur claims to understand what he says and hears, and this because, in a language devoid of convention, the fundamental distinctions between what is certain and what is not, and between what is impossible and what is not, are abolished. Certainty and impossibility then become confused with high or low degrees of a subjective probability, which is itself denied precisely by this falsification of the language.

On the contrary, the preservation of a clear, terse distinction between certainty and uncertainty, impossibility and possibility, is the unique and essential precondition for making *meaningful* statements (which could be either right or wrong), whereas the alternative transforms every sentence into a nonsense.

5.10.6. We have already made abstract reference to this confusion (Section 5.2.3), so let us confine ourselves here to an illustration in the context of physics (with the warning that we are anticipating things to come for the purpose of preventive therapy; later, Chapter 7, we will be concerned with the true meaning of 'laws of large numbers' and suchlike).

It cannot be denied that two different explanations of the same phenomenon may turn out to be indistinguishable in practice; particularly when one explanation is deterministic and the other probabilistic. One thinks immediately of the diffusion of heat, or any other similar phenomenon, which can either be considered in terms of a differential equation, describing the continuous development of the phenomenon in a manner governed precisely by deterministic laws, or as a random process in which elementary phenomena occur in a nondeterministic way, but such that there is a high probability of the phenomenon developing at a macroscopic level in a manner practically identical to that indicated by the deterministic theory.

However, this in no way implies that the two explanations are similar, and even less that they are the same, or substitutable. On the contrary, they are exact opposites; diametrically opposed and absolutely incompatible. The deterministic explanation makes certain assumptions which preclude any departure from predetermined behaviour. Any similar explanation, albeit less rigid, which laid down that some conclusion was compulsory and certain, would, at the very least, require some sort of self-regulatory mechanism, some sort of '*feedback*'. The probabilistic explanation makes no assumptions of this kind: it states nothing other than that everything is possible. If it *appears* to state something more, it is only because such a statement, which may seem quite precise, corresponds to a property common to 'almost all' the possible cases.

A probabilistic explanation of the diffusion of heat *must* take into account the fact that heat could accidentally move from a cold body to a warmer one, making the former even colder, the latter even warmer (in Jeans' example: water being frozen rather than boiled when put on the stove). That this is very improbable is merely due to the fact that the 'unordered' possibilities (heat equally diffused) are far more numerous than the 'ordered' possibilities (all the heat in one direction), and not because the former enjoy some special status.

To rule out the possibility of those cases which seem 'exceptional', in no way *improves* the probabilistic explanation, by somehow making it simpler, or more scientific: on the contrary, it negates it. Acceptance of the probabilistic explanation has the following implications: it means that what we state about the phenomenon must not be regarded as necessary, but, instead, must be attributed to 'chance', and, hence, regarded as only approximate and probable. It means that one must regard it as essential to deny the existence of certain and exact laws which are obeyed only apparently; it means that one must consider as necessary the possibility of studying departures from any rigid law, fluctuations, the effects of discontinuities (the *shot effect*), and all that a cursory identification with a different form of explanation would sweep away without a thought.

5.10.7. What we just said is itself open to misinterpretation. It would be a mistake to infer that an explanation based on a 'tendency to disorder' takes care of every application of probabilistic concepts and not merely the particular example given above.

'Chance' (if we can adopt this convenient terminology as a summary of complicated and uncertain factors without its being taken too seriously) plays a no less important rôle in biological and social processes, where the outcome depends on highly ordered and organized structures, like chromosomes, cells and human beings (and also in physics, in processes like crystallization).

The following needs to be said in order to disprove the thesis which considers a levelling down into a debased chaos (entropic death) to be an inevitable consequence of the validity of this or that 'law' of probability. *The calculus of probability can say absolutely nothing about reality; in the same way as reality, and all sciences concerned with it, can say nothing about the calculus of probability*. The latter is valid whatever use one makes of it, no matter how, no matter where. One can express in terms of it any opinion whatsoever, no matter how 'reasonable' or otherwise, and the consequences will be reasonable, or not, for me, for You, or anyone, according to the reasonableness of the original opinions of the individual using the calculus. As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before (either in terms of having accepted certain facts, or having evaluated degrees of belief in them, respectively).

Physics can make greater or lesser use of the calculus of probability, but the relationship between the two is simply the relationship between a certain field of research, which remains itself, no matter what tools it uses, and a logical tool, unconditionally valid, which remains itself, whatever use is made of it, in whatever field.

5.10.8. Let us return to the necessity of avoiding the dangers implicit in attempts to confuse certainty with 'high probability'. We have to stress this point because these attempts assume many forms and are always dangerous. In one sentence: to make a mistake of this kind leaves one inevitably faced with all sorts of fallacious arguments and contradictions whenever an attempt is made to state, on the basis of probabilistic considerations, that something must occur, or that its occurrence confirms or disproves some probabilistic assumptions.

From such a point of view, the calculus of probability seems to be regarded, more or less explicitly, as a nothingness; saying nothing when the probabilities in question are intermediate in value, but capable of miraculous transformation into a warrantor of absolute truths when the probabilities are very large or very small, since, in these cases, the difference can be ignored and one can simply say that something is true or false. One thus has a mechanism that is considered to be useless when it says that which it is capable of saying, and wishes to say, but is blindly trusted when the things one wants to make it say are not the things it does say or could say.

5.10.9. We present three examples of this form of observation.

First example. The statement that '*an event of small probability does not occur*' is sometimes made, under the heading of 'Cournot's principle' (Section 5.2.3). A kind of corollary or special case of this is referred to as the '*empirical law of chance*' (meaning that frequency and probability actually behave in many cases according to the 'law of large numbers').

Second example. In accordance with the identification of small probability with impossibility, Neyman finds a contradiction in the behaviour of an individual who travels by aeroplane and at the same time takes out insurance. If he considers it possible to

have an accident, why does he travel? If he does not, why does he insure himself? The paradox here specifically relates to 'decision theory', which, in the restricted sense to which it is often reduced by 'objectivist' statisticians, considers only the question 'what decision is appropriate given the *accepted hypothesis*' and not 'what decision is appropriate in the *given state of uncertainty*'.

Third example. Again, in this context (of 'objectivist statisticians'), one aspires to '*accept*' or '*reject*' an hypothesis on the basis of an experiment, instead of considering how its outcome modifies the initial probabilities (which *one wants to do without*!) in order to give the final probabilities (which therefore cannot be obtained!). Here the absurdity reaches new heights, because it cannot even be claimed that 'accept' and 'reject' correspond to the minimal requirement of the probabilities being large or small. The use of these two words is a meaningless convention; an apparent attempt to answer a question by disregarding everything that makes it a meaningful question in the first place.

It is as if, in comparing two weights, we were to decide upon which was the heavier by choosing the one which tilted the balance to its side, without taking into account, and, indeed, refusing to consider it legitimate to take into account, any difference between the arms of the balance, even knowing that the difference could be considerable.

5.11 Determinism, Indeterminism, and other 'Isms'

5.11.1. Continuing with the same theme, there is a clear philosophical point to be made. It derives from the strange fact that precisely the same disposition to accept an objective probability is often justified in two completely opposite ways.

For some people, the ideal instrument for producing an objective probability with value p would be a totally invariable device, working under strictly unchangeable conditions, and for which the tendency to produce successes with frequency p would be a 'built-in property', or, more specifically, a '*dispositional property*' (following, for example, Hacking). Any perturbations would result in a deviation from the desired result; that is, from the realization of a frequency close to p.

For others, whole-hearted determinists, any such device could but yield the same result; always successes or always failures. The fact that both successes and failures occur implies that there exists something causing perturbations. In general, it is assumed that there are a great number of small, accidental, causal factors, which are largely unknown. The fact that the frequency is expected to be around p would be an effect of the combined and random actions of these causal factors (following, for example, Paul Lévy).

So far as the subjectivistic conception is concerned, it has the advantage and, indeed, the preoccupation of remaining outside of disputes of this kind. The thing that really matters, and which justifies, in fact requires, our arguing on the basis of probabilistic logic, is the impossible nature of the situation in which we find ourselves when we attempt to foresee a given outcome with certainty. This is so whatever the reason: whether it be ignorance of certain deterministic laws; or the nonexistence of such laws; or an inability to perform the requisite calculations even though we know the laws; or an inability to obtain precise data (or the impossibility of doing so). At any given time, it does not matter. It is only with respect to the prospects for the advance of science in the future that it matters, and, even here, only in a minor way, since reference to such rigid and preconceived positions seems rather unnecessary anyway.

From time to time, as scientific prospects change, this or that particular mental attitude may be useful, in that it facilitates the formulation of theories which – for the moment – give better agreement with this or that point of view. However, nothing remains for ever unchanged; nothing is absolute. The particular mould in which one sets is not so important: what matters far more is not to set too firmly in any one pattern. To set fast is to no longer be alive.

5.11.2. These same remarks need repeating more generally, in connection with all those ways devised to saddle probability with an objective something (meaning, interpretation, justification, definition or whatever). In the first place, it is a fact that these attempts are not successful, and cannot be so, since, having the resolve to express matters relating to uncertainty in terms of the logic of certainty, they force themselves, *ab initio*, into a vicious circle, with no means of escape – 'per la contradizion che no'l consente'.[24] It is as if someone were to wish to hoist himself up by his bootlaces. Logic only permits the exposure of a tautology on the basis of what is taken as known; a prevision, however, is not simply a tautological consequence of what is already known. To be thus, would be to constitute something implicitly known, and would not involve uncertainty and, therefore, would not give rise to prevision.

However, even if we were to consider someone's arguments to constitute an acceptable basis for an objective meaning of probability (and, in general, such arguments will be different and concerned with special and different types of event, according to the different points of view), our thesis consists in believing that these arguments would be irrelevant anyway. All such conceptions, all the 'isms' they reduce to, are rejected here, but *not* in support of yet another 'ism' (as might be thought; e.g. 'subjectivism' or 'solipsism'), which one wants to put forward and contrapose to the others. The latter are rejected because, whatever the explanation of the uncertainty might be (attributing it to 'chance', 'fate', 'hidden laws', 'Providence', or 'statistical regularities', or to something else – or … words(?)…), the sole concrete fact which is beyond dispute is that someone (me, You, somebody else) feels himself in a state of uncertainty, and has to decide on and adopt some point of view as a basis for previsions and related decisions.

5.11.3. This subjective meaning is an objective and unquestionable fact: all the rest (even if there were no dispute about it) is, in any case, something of an extra, which, at best, serves to help fix one's ideas. It is analogous to a vivid piece of writing that succeeds in forming something like an idea in our minds, although its meaning is not clear, and an analysis of the sentence in fact shows it to be inconsistent.

It is the case, however, that this view of the logic of uncertainty, complete and clear as it is, is far from achieving general acceptance. Why is this? Perhaps it is only the state of being certain which appears to most people as worthy of consideration and fit to be part of the edifice of science (which, according to the prevailing view, appears to express or aspire to omniscience – notwithstanding the fact that all progress, pushing back as it does the frontier of what is known, makes the horizon of what can be seen as unknown even broader). Perhaps the unknown and the uncertain disturb and annoy us, and

24 *Translators' note*. Ruled out by the principle of contradiction. (The line is from Dante's Inferno: canto XXVII, line 120.)

provoke those who are most upset to attempt to suppress them, or at least to make them disappear. There is not much point in philosophical arguments or speculations of this kind: they do provide, however, a possible explanation of why these different attitudes exist (we had to mention them, or, at least, to give some indication of how a person who finds our point of view natural could try to understand its lack of acceptance).

These different attitudes are, essentially, only variations on the same theme: the attempt to avoid the problems of uncertainty by simply pretending to overcome it; restricting the treatment to cases in which it can be presented in such a watered down way that it looks like something else.

The classical variant limits itself to cases like those of games of chance (where probability should acquire an objective meaning by virtue of the 'definition' based upon 'equally likely cases'). In the view of the most rigid supporters of this position, every application of the theory of probability outside this field would only be a questionable transposition by analogy.

The position that is at present most widely accepted restricts itself to cases of a certain statistical type (where probability should acquire an objective meaning by virtue of the 'definition' based upon 'frequency'). According to its most rigid adherents, the term 'probability', when used outside this context, has no more in common with its 'scientific' meaning than the 'energy' of a team leader has with the same term as applied to physical motion.[25]

Other approaches, which, having the aim of acting as guides in decision making, follow less rigid notions, attempt nevertheless to avoid those components of the argument which many find unpalatable (like the 'initial probabilities' required for Bayesian induction).[26]

Others adopt an eclectic attitude, accepting that one can base one's thinking on 'that probability which we evaluate for previsions and decisions' (i.e. the one corresponding to the conception of the present author), but, on the other hand, asserting that 'there is also another type of probability, the one with which statistics is concerned' (or, alternatively, 'the type valid in games of chance', or both).[27]

We should point out, here and now, that the mathematical treatment is unaffected (or, at most, very little affected) by these disagreements. In this sense, we can give a reassurance that everything we shall say mathematically is independent of questions of this kind, and should be acceptable to everyone. However, the interpretation is often different; there are certain nuances which, when looked into closely, completely change the spirit in which a given statement (perhaps expressible, in the same words, in the imprecise manner of everyday language) is to be understood.

So far as our own attitude is concerned, we wish to make clear that it is not utterly opposed to the attitude we have termed 'eclectic', even though it differs from it in a very real sense.

It is not utterly opposed because we recognize the importance of the problems, concepts and criteria that are the object of the various practical theories, even though we

---

25 The phrases given here, in characterizing the two attitudes, are due to Castelnuovo (transposition by 'analogy') and von Mises ('energy' and energy), respectively.

26 The followers of 'objectivistic statistics' in its various schools, including that of A. Wald (who we particularly have in mind here, as the nearest in approach to the Bayesian school).

27 The quotations are from V. Castellano. Typical examples of the eclectic attitude are provided by R. Carnap (who differentiates between 'probability$_1$' *logical*, and 'probability$_2$', *statistical*) and I.J. Good who admits the possible value of distinguishing many 'kinds of probability' (although in the context of a conception which is essentially subjectivistic).

study them within the framework of the general theory. Only by renouncing their alleged autonomy is it possible to compensate for those deficiencies in the foundations of the particularistic theories which render their conclusions meaningless, and the interpretation of them arbitrary.

It differs from it because we do not accept the existence of probabilities of different kinds, nor the autonomous validity of theories which set out to consider them, leaving aside some of the assumptions of the general theory, all of which are at all times essential.

All this has been summed up in an expressive manner by L.J. Savage (in a rather more specialized context): it is as though one wished to make a probabilistic omelette without breaking probabilistic eggs. There are two possible outcomes: either the result is not an omelette; or the eggs have in fact been used, either surreptitiously or inadvertently. All comments that we shall have occasion to make concerning 'other points of view' will essentially be continuations of the above analogy.

6

Distributions

6.1 Introductory Remarks

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

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

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

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

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

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

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

6.2 What we Mean by a 'Distribution'

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

where:

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

We now consider the residual partial distribution,

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

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

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

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

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

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

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

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

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

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

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

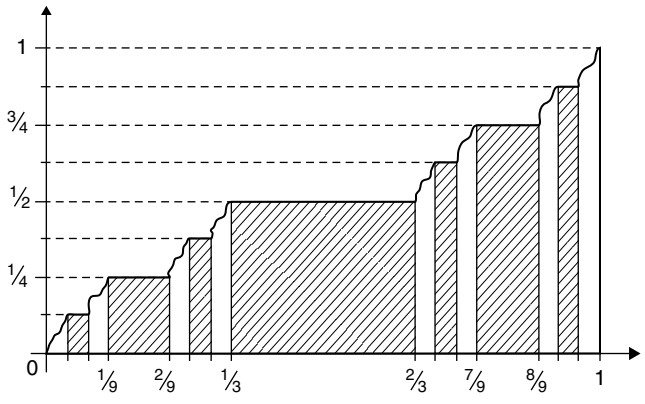


Figure 6.1 The Cantor distribution.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As an example, suppose we consider the two functions

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

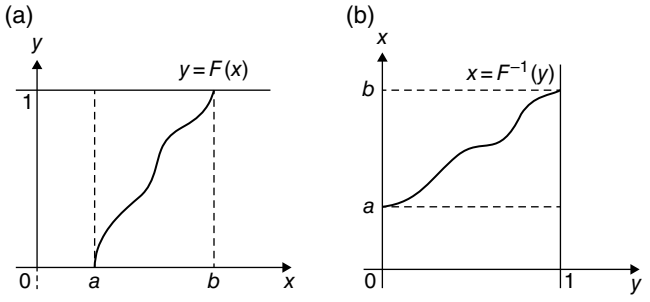


Figure 6.2 The graphs of (a) the distribution function and (b) its inverse: $y = F(x)$ and $x = F^{-1}(y)$.
In addition to the present (measure theoretic) interpretation, we have also seen that the statistical interpretation, as *graph of the distribution*, is of interest (and is the most useful from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

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

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

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

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

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

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

6.3 The Parting of the Ways

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

6.4 Distributions in Probability Theory

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.5 An Equivalent Formulation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

or, in a different but equivalent form,

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

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

For X non-negative and bounded, we certainly have

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

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

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

this holds for any K, and hence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.6 The Practical Study of Distribution Functions

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

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

descriptive properties,
synthetic characteristics,
analytic characteristics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

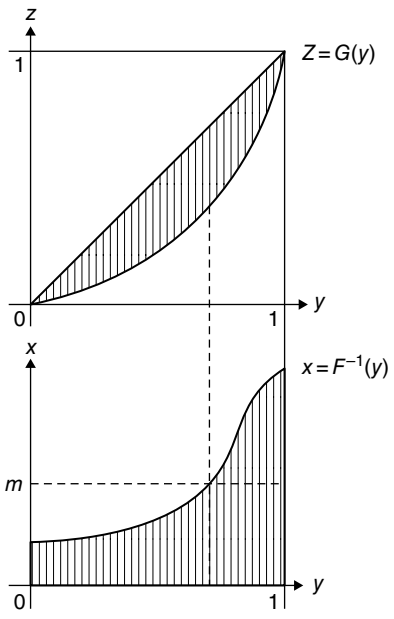


Figure 6.3 The concentration curve $z = G(y)$; for example, in the case of incomes, to the fraction y of the least wealthy, there corresponds the fraction $G(y)$ of total income, which is represented on the graph below (the gradation curve: see Figure 6.2b and the discussion in 6.2.6) by the fraction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

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

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

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

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

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

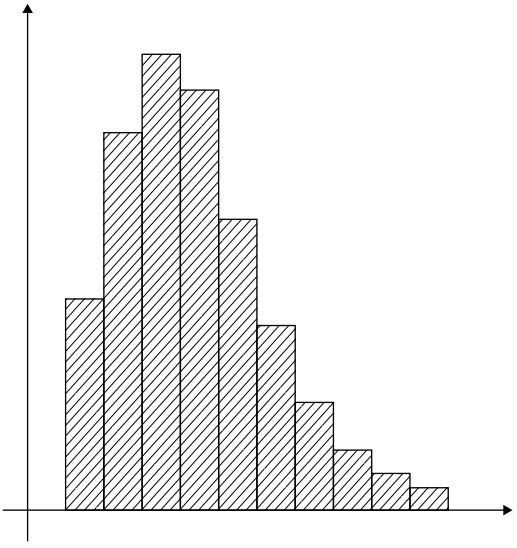


Figure 6.4 An example of a histogram. (It represents the distribution of families in Italy in 1951, according to the number in each.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

6.7 Limits of Distributions

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

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

given any $\varepsilon > 0$, the inequalities

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

are satisfied for all n greater than some N.

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

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

6.7.2. We can straightaway make some important points.

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

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

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

---

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

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

or, respectively,

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

from which it follows that

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

As a result :

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

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

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

6.8 Various Notions of Convergence for Random Quantities

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

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

---

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

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

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

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

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

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

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

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

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

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

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

We give straightaway the three most important types of convergence.

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

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

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

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

  or

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

stated explicitly,

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

and, more generally, we have the similar result

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

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

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

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

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

Definite uniform convergence implies all the above.

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

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

---

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

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

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

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

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

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

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

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

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

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

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

In particular, it could in this case be

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

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

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

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

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

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

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

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

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

6.9 Distributions in Two (or More) Dimensions

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

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

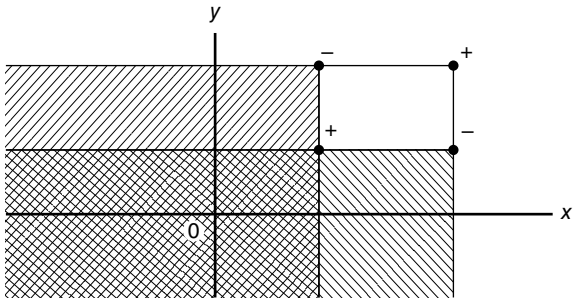


Figure 6.5 Quadrants of the (x, y)-plane, in terms of which the joint distribution function $F(x, y)$ is defined (SW quadrants), and a method of indicating the rectangles with their linear combinations (and, hence, their probabilities in terms of linear combinations of the values $F(x, y)$ at the vertices).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

And it would be easy to continue in this manner.

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

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

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

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

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

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

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

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

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

shall have

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

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

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

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

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

For the examples we have considered, this gives

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

is also equivalent to the following condition:

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

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

in other words,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

For the third-order moments, we have

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

and, in the case of independence,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

so that, in the case of independence,

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

We shall see shortly how this property can be exploited.

6.10 The Method of Characteristic Functions

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

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

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

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

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

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

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

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

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

we have an 'automatic' way of computing the probabilities

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

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

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

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

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

This justifies the name *characteristic function* given to

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

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

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

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

or

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

Formula 6.31 is the classical Fourier *inversion theorem*.

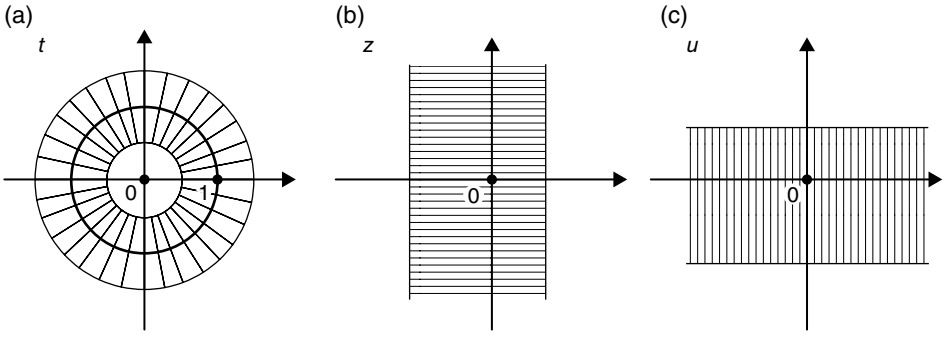


Figure 6.6 The planes of the three variables *t*, *z* and *u*, in terms of which the characteristic function can be expressed, together with the lines or regions where it is defined. Usually we operate in terms of *u* (the Fourier transform); $z = iu$ and $t = e^z = e^{iu}$ are occasionally to be preferred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

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

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

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

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

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

---

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

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

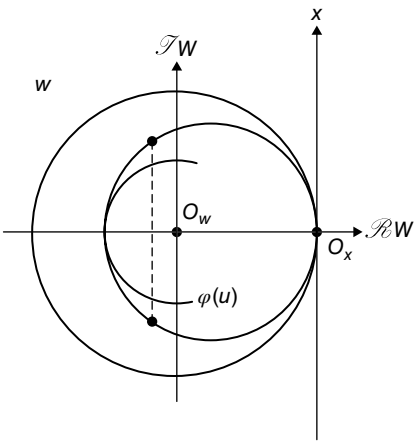


Figure 6.7 The plane of $w = \phi(u)$, and the interpretation of $\phi(u)$ as the barycentre of the distribution 'wrapped around the circumference $|w| = 1$'.

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

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

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

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

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

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

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

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

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

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

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

When the densities exist, we have

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

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

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

---

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

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

6.11 Some Examples of Characteristic Functions

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

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

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

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

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

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

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

whereas

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and that

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

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

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

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

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

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

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

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

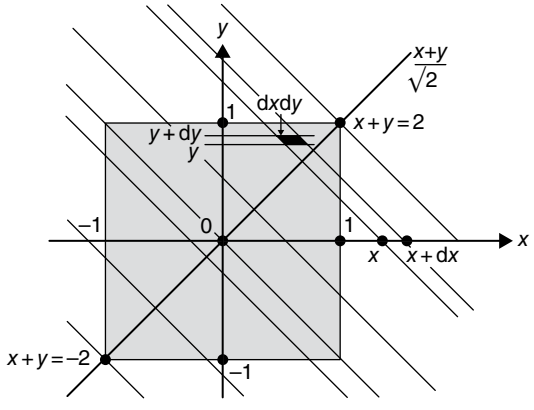


Figure 6.8 The convolution of uniform distribution.

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

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

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

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

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

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

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

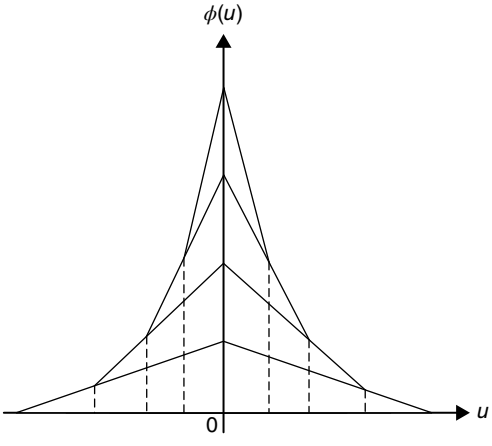


Figure 6.9 Characteristic functions constructed on the basis of Pólya's argument.

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

For the *exponential distribution*,

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

and

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

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

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

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

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

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

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

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

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

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

By inversion, we obtain

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

or even

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.12 Some Remarks Concerning the Divisibility of Distributions

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

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

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

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

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

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

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

7

A Preliminary Survey

7.1 Why a Survey at this Stage?

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

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

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

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

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

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

---

1 Feller, for example, repeatedly remarks on the way in which certain results seem to be surprising and even paradoxical (even simple results concerning coin tossing, such as those relating to the periods during which one gambler has an advantage over the other; Chapter 8, 8.7.6). He remarks on 'conclusions that play havoc with our intuition' (p. 68), and that 'few people will believe that a perfect coin will produce preposterous sequences in which no change of lead occurs for millions of trials in succession, and yet this is what a good coin will do rather regularly' (p. 81). Moreover, he can attest to the fact that 'sampling of expert opinion has revealed that even trained statisticians feel that' certain data are really surprising (p. 85).
All this is from W. Feller, *An Introduction to Probability Theory and its Applications*, 2nd edn, John Wiley & Sons, Inc., New York (1957). Many of the topics mentioned in the present chapter are discussed in detail by Feller (in Chapter 3, in particular), who includes a number of original contributions of his own.

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

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

7.2 Heads and Tails: Preliminary Considerations

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

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

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

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

---

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

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

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

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

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

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

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

We shall return to this topic again in Section 7.6.9.

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

For runs of length *m*, we shall give, for simplicity, the asymptotic expression for *n* large in comparison to *m* (see (*L*)). For each toss (and, to be rigorous, we should modify

this for the initial and final m tosses), we have probability $m/2^{m+1}$ of belonging to a run of length m, and so, in prevision, there are $nm/2^{m+1}$ such tosses out of n; there are, therefore, $n/2^{m+1}$ runs of this kind (since each consists of m tosses). In particular, we have, in prevision, $n/4$ isolated outcomes, $n/8$ runs of length two, and so on.

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

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

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

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

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

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

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

7.3 Heads and Tails: The Random Process

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

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

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

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

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

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

---

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

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

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

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

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

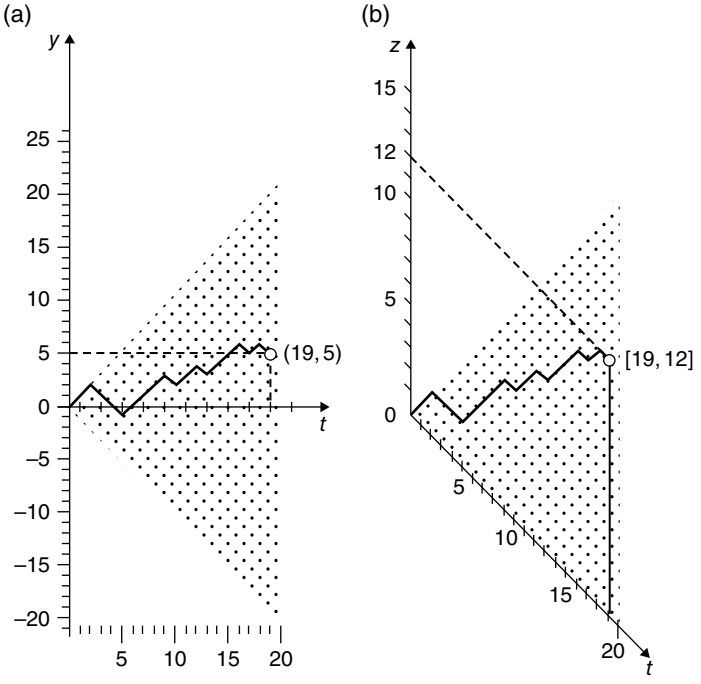


Figure 7.1 The lattice of possible points for the coin-tossing process. Coordinates: time $t = n$ = number of tosses (in both cases), together with: in (a): y = gain = $h - (n - h)$ = number of successes minus number of failures; in (b): $z = h$ = number of successes = $(n + y)/2$. The notations (t,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

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

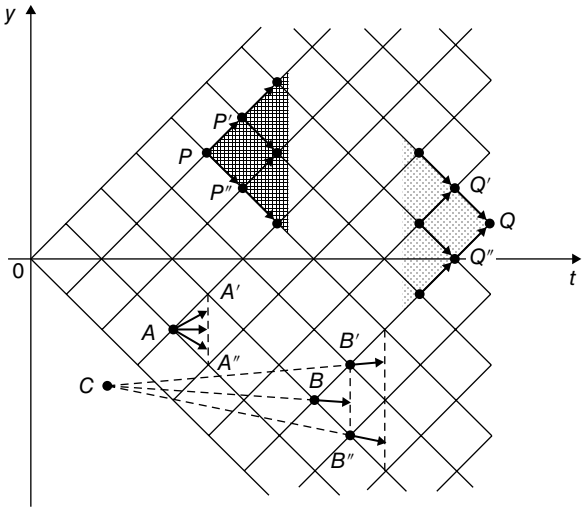


Figure 7.2 The lattice of the 'random walk' for coin tossing (and similar examples). The point Q can only be reached from Q' or Q'' (this trivial observation often provides the key to the formulation of problems). It follows, therefore, that it can only be reached from within the angular region shown.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

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

movements are, the mass is alternatively all in the even positions or all in the odd: in any case, we have some kind of *diffusion* process (and this is more than just an image). In many problems, it happens, for example, that under certain conditions the process comes to a halt, and this might correspond to the mass being stopped by coming up against an absorbing barrier.

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

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

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

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

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

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

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

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

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

Successes in n tosses. We already know (case (A), Section 7.2.2) that the probability of h successes in n tosses is given by

$$\omega_h^{(n)} = \binom{n}{h}/2^n \quad (h = 0, 1, 2, \ldots, n). \tag{7.1}$$

14 Figure 7.3 shows the results of successive subdivisions, $(\frac{1}{2},\frac{1}{2})$,$(\frac{1}{4},\frac{1}{2},\frac{1}{4})$,$(\frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8})$ etc. Figure 7.4 shows a simple apparatus invented by Bittering. It is a box with two sets of divisions into compartments, one set being on top of the other and shifted through half the width of a compartment. The middle section of the bottom half of the box is filled with sand and then the box is turned upside down. The sand now divides itself between the two central compartments of what was the top half and is now the bottom. By repeatedly turning the box over (shaking it each time to ensure a uniform distribution of the sand within the compartments), one obtains successive subdivisions (the ones we have referred to above – those of Heads and Tails). By arranging the relative displacement of the overlapping compartments to be in the ratio p:1 – p, one can obtain any required Bernoulli distribution (see Section 7.4.2).

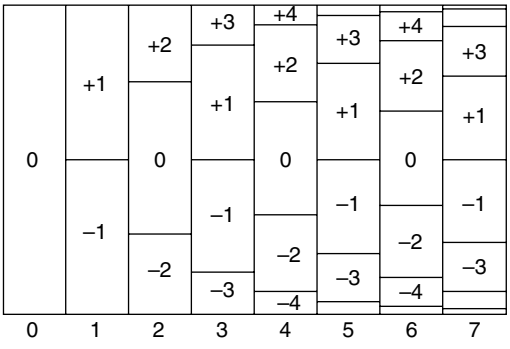


Figure 7.3 Subdivision of the probability for the game of Heads and Tails.

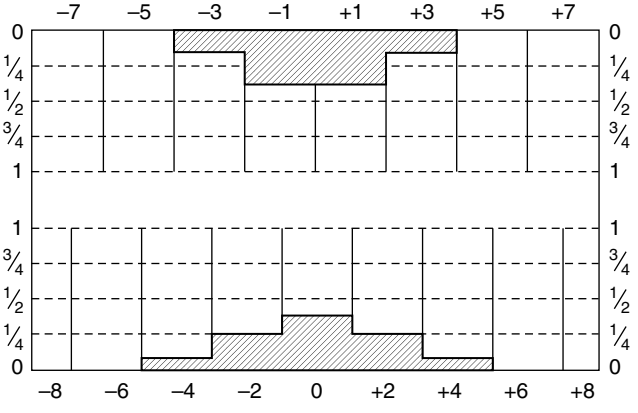


Figure 7.4 Bittering's apparatus. Probabilities of Heads and Tails: below, after four tosses; above, after three tosses.

This corresponds to the fact that $\binom{n}{h}$ is the number of paths which lead from the origin 0 to the point $[n, h]$.[15]

To see this, consider, at each point of the lattice, the number of paths coming from 0. This number is obtained by summing the numbers corresponding to the two points adjacent on the left, since all relevant paths must pass through one or other of these. 'Stiefel's identity', $\binom{n-1}{h-1}+\binom{n-1}{h}=\binom{n}{h}$, provides the key and leads one to the binomial coefficients of 'Pascal's triangle'. That the total number of paths is 2^n follows directly from the fact that at each step each path has precisely two possible continuations.

The identity which we mentioned in example (D) of Section 7.2.2 also finds an immediate application. Each of the $\binom{N}{H}$ jagged lines which lead to a given point $[N, H]$ must pass through the vertical at n at some point $[n, h]$, where, since n is less than N, $h = S_n$ must necessarily satisfy

$$H+n-N \leqslant h \leqslant H$$

15 Recall that, in our notation, $[n, h]$ represents the fact that $S_n = h$; in other words, it represents the point $(n, 2h - n)$, where $Y_n = 2S_n - n = 2h - n$.

(because, between n and N, there can be no reduction in either the number of successes or of failures). There are $\binom{n}{h}$ paths from the origin arriving at $[n, h]$, and $\binom{N-n}{H-h}$ paths leading from this point to $[N, H]$. There are, therefore, $\binom{n}{h}\binom{N-n}{H-h}$[16] paths from the origin to $[N, H]$ which pass through the given intermediate point $[n, h]$; summing over the appropriate values of h, we must obtain $\binom{N}{H}$, thus establishing the identity. Further, we see that

$$\binom{n}{h}\binom{N-n}{H-h}/\binom{N}{H}$$

is the probability of passing through the given intermediate point conditional on arriving at the given final destination (in other words, we obtain

$$\mathbf{P}\left(Y_n = 2h-n \,|\, Y_N = 2H-N\right) \quad \text{or} \quad \mathbf{P}(S_n = h \,|\, S_N = H);$$

we shall return to this later).

The rth *success.* Problem (F) of 7.2.4 can be tackled in a similar fashion, by reasoning in terms of crossings of sloping lines rather than vertical ones. In fact, the rth success is represented by the rth step upward; that is, the step which takes one from the line $y = 2r - 2 - t$ to the line $y = 2r - t$ (i.e. from the rth to the $(r + 1)$th downward sloping line of the lattice, starting from $y = -t$). It is obvious that the rth failure can be dealt with by simply referring to upward sloping lines rather than downward ones. We have already shown the probability of the rth success at the hth toss to be $\binom{h-1}{r-1}/2^h$. We note that the favourable paths are those from 0 to $[h, r]$ whose final step is upward, that is passing through $[h - 1, r - 1]$, and that there are, in fact, $\binom{h-1}{r-1}$ of these.

If one is interested in considering the problem conditional on the path terminating at $[N, H]$, it is easily seen that there are $\binom{h-1}{r-1}\binom{N-h}{H-r}$ paths in which the rth success occurs at the hth toss (they must go from 0 to $[h - 1, r - 1]$, and then, with a compulsory step, to $[h, r]$, and finally on to $[N, H]$). As above, we can sum over all possibilities to obtain $\binom{n}{h}$[17] (the sum being taken over $r \leqslant h \leqslant N - H + r$, since r successes cannot occur until at least r trials have been made, nor can there be more than $H - r$ failures in the final $H - r$ trials).

Dividing the sum by the total, we obtain, in this case also, the conditional probabilities (of the rth success at the hth toss, given that out of N tosses there are H successes; we must have $r \leqslant h \leqslant H\,N - H + r$). These are given by

---

16 It is often not sufficiently emphasized that the *basic operation of combinatorial calculus is the product*; this should always be borne in mind, using this and many similar applications as examples.

17 In this way, we arrive at meaningful interpretations of two well-known identities involving products of binomial coefficients:

$\binom{N}{H} = \sum_{h=0\vee(H+n-N)}^{n\wedge H} \binom{n}{h}\binom{N-n}{H-h}$ (holding for each fixed n, with $\mathbf{1} \leqslant n \leqslant N$).

$\binom{N}{H} = \sum_{h=r}^{N-H+r} \binom{h-1}{r-1}\binom{N-h}{H-r}$ (holding for each fixed $\boldsymbol{r}$, with $1 \leqslant r \leqslant H$).

These simply give the number of paths from 0 to $[N, H]$, expressed in terms of the points at which they cross vertical (1st identity) or sloping (2nd identity) lines.

$$\binom{h-1}{r-1}\binom{N-h}{H-r}\Big/\binom{N}{H} = \mathbf{P}\left(S_{h-1}+1=S_h=r \mid S_N=H\right)$$
$$= \mathbf{P}\left(Y_{h-1}+1=Y_h=2r-h \mid Y_N=2H-N\right).$$

This result will also be referred to again later.

Gambler's ruin. The problem of the crossings of horizontal lines is more complicated, as, unlike the previous cases, more than one crossing is possible (in general, an unlimited number). It is, however, a very meaningful and important topic, and, in particular, relates to the classical gambler's ruin problem.

If a gambler has initial fortune c, then his ruin corresponds to his gain reaching $-c$. Similar considerations apply if two gamblers with limited fortunes play against each other. Here, we confine ourselves to just this brief comment but we note that, for this and for other similar problems (some of them important), arguments in terms of paths will turn out to be useful. In particular, we shall make use of appropriate *symmetries* of paths by means of Desiré André's celebrated *reflection principle* (in particular, Chapter 8 will deal with topics of this kind).

7.4 Some Particular Distributions

7.4.1. Before we actually begin our study of random processes, we shall, on the basis of our preliminary discussions, take the opportunity to examine a few simple problems in more detail, and to consider some particular distributions.

In order to avoid repetition later (and for greater effectiveness), we shall consider these distributions straightaway, both in the special forms that are appropriate for Heads and Tails, and in the more general forms. It should be noted that although the form of representation which we have adopted is a valid and useful one, the property of *fairness* (together with the principle of *reflection* and the *equal* probabilities of the paths) only holds for the special case of Heads and Tails.

In order to achieve some uniformity in notation, we shall always use X to denote the random quantity under consideration, and $p_h = \mathbf{P}(X = x_h)$ to denote the probability concentrated at the point x_h.[18] In the examples we shall consider, however, it turns out that the possible values of x_h are always integer (apart from changes in scale, $x_h = h$). For the particular case of the 'number of successes', we shall always use $\omega_h^{(n)} = \mathbf{P}(S_n = h)$ for the p_h.

7.4.2. *The Bernoulli* (*or binomial*) *distribution*. This is the distribution of $S_n = E_1 + E_2 + \ldots + E_n$ (or of $Y_n = 2S_n - n$, or of the frequency S_n/n – they are identical apart from an irrelevant change of scale) when the events E_h are *independent and have equal probabilities*, $\mathbf{P}(E_h) = p$. When $p = \frac{1}{2}$, as in the case of Heads and Tails, we have the *symmetric* Bernoulli distribution. The distributions are, of course, different for different n and p. Given n, the possible values are $x_h = h = 0, 1, 2, \ldots, n$ (or $x_h = a + hb = a, a + b, a + 2b, \ldots, a + nb$), and their probabilities are given by

$$p_h = \binom{n}{h} p^h \tilde{p}^{n-h} \quad \left(\text{if } p = \frac{1}{2},\ p_h = \binom{n}{h}/2^n\right) \tag{7.2}$$

that is the $\omega_h^{(n)}$ of the process.

18 We shall use *concentrated* rather than *adherent* (see Chapter 6) because, in these problems, the possible values can only be, by definition, the x_h themselves (finite in number, and, in any case, discrete).

For the case $p = \frac{1}{2}$, we know that $\mathbf{P}(X) = n/2$, and $\boldsymbol{\sigma}(X) = \sqrt{n}/2$ (see Section 7.3.4). Similarly, for arbitrary p, we see that $\mathbf{P}(X) = np$, $\boldsymbol{\sigma}(X) = \sqrt{(np\tilde{p})}$, because for each summand we have

$$\mathbf{P}(E_i) = \mathbf{P}(E_i^2) = p, \qquad \sigma^2(E_i) = \mathbf{P}(E_i^2) - \mathbf{P}^2(E_i) = p - p^2 = p\tilde{p}.$$

Hence, using the second-order properties, we obtain, without calculation,

$$\begin{aligned} &\sigma^2(X) = \sum_{h=0}^{n} \binom{n}{h} p^h \tilde{p}^{n-h} (h - np)^2 = np\tilde{p} \\ &\text{(in addition to } \mathbf{P}(X) = \sum_{h=0}^{n} \binom{n}{h} p^h \tilde{p}^{n-h} h = np). \end{aligned} \tag{7.3}$$

The behaviour of the $p_h = \omega_h^{(n)}$ in the case $p = \frac{1}{2}$ is governed by that of the binomial coefficients $\binom{n}{h}$. These are largest for central values ($h \simeq n/2$) and decrease rapidly as one moves away on either side. Unless one looks more carefully[19] at ratios like p_{h+1}/p_h, however, it is difficult to get an idea of *how rapidly* they die away:

$$\frac{\omega_{h+1}^{(n)}}{\omega_h^{(n)}} = \frac{n-h}{h+1} \quad \left(\text{in general, } \frac{n-h}{h+1} \cdot \frac{p}{\tilde{p}} \text{ for } p \neq \frac{1}{2} \right). \tag{7.4}$$

The same conclusions hold for general p, except that the maximum is attained for some $h \simeq np$ (instead of $\simeq \frac{1}{2} n$).

In fact, a consideration of Tchebychev's inequality suffices to show that the probability of obtaining values far away from the prevision is very small. Those h which differ from np by more than $n\varepsilon$ ($\varepsilon > 0$), that is corresponding to frequencies h/n not lying within $p \pm \varepsilon$, have, *in total*, a probability less than $\boldsymbol{\sigma}^2(X)/(n\varepsilon)^2 = np\tilde{p}/(n\varepsilon)^2 = p\tilde{p}/n\varepsilon^2$ (and this is far from being an accurate bound, as will be clear from the asymptotic evaluations which we shall come across shortly; equation 7.20 of Section 7.5.4).

Comments. The p_h can be obtained as the coefficients of the expansion of

$$(\tilde{p} + pt)^n = \sum_{h=0}^{n} p_h t^h = \sum_{h=0}^{n} \omega_h^{(n)} t^h$$

(the alternative notation being chosen to avoid any ambiguity in the discussion which follows). It suffices to observe that the random quantity

$$\prod_{i=1}^{n} (\tilde{E}_i + tE_i)$$

is the sum of the constituents multiplied by t^h, where h is the number of positive outcomes (giving, therefore, $S_n = h$). Its value is thus given by

$$\sum_{h=0}^{n} (S_n = h) t^h = t^{S_n},$$

19 As is done, for the purpose of providing an elementary exposition, in B. de Finetti and F. Minisola, *La matematica per le applicazioni economiche*, Chapter 4. See also a brief comment later (in Section 7.6.3).

and its prevision, that is the characteristic function $\phi(u)$ with $t = e^{iu}$ (or $u = -i \log t$), by

$$\mathbf{P}\left(t^{Sn}\right) = \sum \mathbf{P}\left(S_n = h\right) t^h. \tag{7.5}$$

A generalization. In the same way, we observe that the result holds even if the E_i (always assumed stochastically independent) have different probabilities p_i. In this case, the $\omega_h^{(n)}$ are given by

$$\sum_{h=0}^{n} \omega_h^{(n)} t^h = \prod_{i=1}^{n} \left(\tilde{p}_i + p_i t\right). \tag{7.6}$$

On the other hand, this only expresses the obvious fact that $\omega_h^{(n)}$ is the sum of the products of h factors involving the p_i, and $n - h$ involving the complements $\tilde{p}_i$.

In particular, we see that $\sigma^2\left(S_n\right) = \sum_i p_i \tilde{p}_i$, and that this formula (like $np\tilde{p}$, of which it is an obvious generalization) continues to hold, even if we only have pairwise independence. (Recall that this is not sufficient for many other results concerning the distribution.)

7.4.3. *The hypergeometric distribution*. As in the previous case, we are interested in the distribution of $X = S_n$ (or $Y_n = 2S_n - n$, or S_n/n, which, as we have already remarked, only differ in scale). The difference is that we now *condition on the hypothesis that, for some given $N > n$, we have $S_N = H$.*

In deriving the required distribution, it suffices that the $\binom{N}{H}$ paths from 0 to $[N, H]$ appear equally probable to us. It does not matter, therefore, whether we choose to think in terms of Heads and Tails (where initially all 2^N paths were equally probable, and the paths compatible with the hypothesis remain such), or in terms of events which, prior to the hypothesis, were judged independent and equally probable, but with $p \neq \frac{1}{2}$ (because in the latter case all the remaining $\binom{N}{H}$ paths have the same probability, $p^H \tilde{p}^{N-H}$).

Instead of thinking in terms of these representations (whose main merit is that they show the links with what has gone before), it is useful to be able to refer to something rather more directly relevant. The following are suitable examples: drawings without replacement from an urn (containing N balls, H of which are white); counting votes (where a total of N have been cast, H of which are in favour of some given candidate); ordering N objects, H of which are of a given kind (N playing cards, H of which are 'Hearts'; N contestants, H of whom are female). In all these cases, the $N!$ possible permutations are all considered equally probable. (Or, at least, all $\binom{N}{H}$ possible ways of arranging the two different kinds of objects must be regarded as equally probable; it is these which correspond to the $\binom{N}{H}$ paths involving H upward steps and $N - H$ downward steps.)

Under the given assumptions (or information), each event $E_i (i \leqslant N)$ has probability $\mathbf{P}(E_i) = H/N$[20] (and, for convenience, we shall write $H/N = q$). These events are not independent; in fact, we shall see later that they are negatively correlated.

---

20 Given that we assume the hypothesis $S_N = H$ to be already part of our knowledge or information, we take $\mathbf{P}(E)$ to mean $\mathbf{P}(E|S_N = H)$. In this situation, the E_i have probability $q = H/N$, but are not stochastically independent (even if they are the outcomes of coin tossing, or rolling a die etc., where, prior to the information about the frequency of successes out of N tosses, they were judged independent and equally probable). In particular, $p_h = \omega_h^{(n)} = \mathbf{P}(S_n = h)$ *in this case* is what we would have written as $\mathbf{P}(S_n = h|S_N = H)$ in the previous case.

Observe that, as a result of changes in the state of information, problems which were initially distinct may come to be regarded as identical, and assumptions about equal probabilities, or independence, may cease to hold (or conversely in some cases). These and other considerations will appear obvious to those who have entered into the spirit of our approach. Those who have come to believe (either through ignorance or misunderstanding) that properties like stochastic independence have an objective and absolute meaning that is inherent in the phenomena themselves, will undoubtedly find these things rather strange and mystifying.

The distribution that concerns us (for example, that of the number of white balls appearing in the first n drawings – or something equivalent in one of the other examples) will be different for every triple n, N and H (or, equivalently, n, N, q). For $q = \frac{1}{2}$, that is for $H = N - H = \frac{1}{2}N$, the distribution is symmetric; $\mathbf{P}(S_n = h) = \mathbf{P}(S_n = n - h)$. The possible values are the integers $x_h = h$, where $0 \vee H - (N - n) \leqslant h \leqslant n \wedge H$ (or $x_h = a + hb$; e.g. $= 2h - n$, or $= h/n$) and their probabilities are, as we saw already in 7.3.4,

$$
\begin{aligned}
p_h = \omega_h^{(n)} &= \frac{\binom{n}{h}\binom{N-n}{H-h}}{\binom{N}{H}} \\
&= \frac{\binom{H}{h}\binom{N-H}{n-h}}{\binom{H}{n}} \\
&= \binom{n}{h} \frac{\begin{array}{c} H(H-1)(H-2)\ldots(H-h+1)(N-H)(N-H-1)(N-H-2) \\ \ldots(N-H-(n-h)+1) \end{array}}{N(N-1)(N-2)\ldots(N-n+1)} \\
&\qquad \frac{\left[\left(1-\frac{1}{H}\right)\left(1-\frac{2}{H}\right)\ldots\left(1-\frac{h-1}{H}\right)\right]}{} \\
&= \binom{n}{h} q^h \tilde{q}^{n-h} \frac{\left[\left(1-\frac{1}{N-H}\right)\left(1-\frac{2}{N-H}\right)\ldots\left(1-\frac{n-h-1}{N-H}\right)\right]}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)}.
\end{aligned}
\tag{7.7}
$$

The interpretation of the four different forms is as follows.

The *first form* (as we already know) enumerates the paths.

The *second form* enumerates those n-tuples, out of the total of $\binom{N}{H}$ that can be drawn from N events, which contain h out of the H, and $n - h$ out of the $N - H$.

The *third form* (which can be derived from the previous two) can be interpreted directly, observing that the probability of first obtaining h successes, and then $n - h$ failures, is given by the product of the ratios (of white balls, and then of black balls) remaining prior to each drawing:

$$
\begin{aligned}
&\frac{H}{N}\cdot\frac{H-1}{N-1}\cdot\frac{H-2}{N-2}\cdots\frac{H-h+1}{N-h+1}\cdot\frac{N-H}{N-h}\cdot\frac{N-H-1}{N-h-1} \\
&\qquad\qquad \times\frac{N-H-2}{N-h-2}\cdots\frac{N-H-(n+h)+1}{N-n+1}.
\end{aligned}
$$

But this is also the probability for any other of the $\binom{n}{h}$ orders of drawing this number of successes and failures (even if the ratios at each drawing vary, the result merely involves permuting the factors in the numerator, and is, therefore, always the same).

We see already that, provided n is small in comparison to N, H and $N - H$, all the ratios differ little from q (the drawings already made do not seriously alter the composition of the urn). The results will, therefore, not differ much from the Bernoulli case (drawings from the same urn *with* replacement).

The *fourth form* shows the relation between the two cases explicitly, by displaying the correction factor.

The behaviour of the p_h in this case is similar to that of the Bernoulli case, and can be studied in the same way (by considering ratios p_{h+1}/p_h). The maximum is obtained by the largest h which does not exceed

$$nq\left[1-2/\left(N+2\right)\right]-\left(H-3\right)/\left(N+2\right)$$

(the reader should verify this!), and as one moves further and further away on each side, the p_h decrease. Compared with the Bernoulli case, the terms around the maximum are larger, and those far away are smaller. Some insight into this can be obtained by looking at the final formula.[21]

For the prevision, we have, of course, $\mathbf{P}(X) = nq = nH/N$. The standard deviation $\boldsymbol{\sigma}(X)$, on the other hand, is a little smaller than $\sqrt{(npp\tilde{p})}$ (the result for the case of independence), and is given by

$$\sigma^2\left(X\right)=nq\tilde{q}\left[1-\left(n-1\right)/\left(N-1\right)\right].$$

In fact, if we evaluate the correlation coefficient r between two events ($r = r(E_i, E_j)$, $i \neq j$) we obtain $r = -1/(N - 1)$. More specifically,

$$\mathbf{P}\left(E_iE_j\right)=\left(H/N\right)\left(\left(H-1\right)/\left(N-1\right)\right)=q^2\left(1-1/H\right)/\left(1-1/N\right),$$

from which it follows that

$$\begin{aligned} r&=\left[\mathbf{P}\left(E_iE_j\right)-\mathbf{P}\left(E_i\right)\mathbf{P}\left(E_j\right)\right]/\sigma\left(E_i\right)\sigma\left(E_j\right)\\ &=q^2\left[\left(N+H-1\right)/N\left(N-1\right)\right]/q\tilde{q}=-1/\left(N-1\right).\end{aligned}$$

---

21 For this correction factor, the variant of Stirling's formula (equation 7.28) which is given in equation 30 (see 7.6.4) yields the approximation (for $n \ll N$)

$$\exp\left\{-\frac{1}{2}\frac{n}{N}\left[1+\frac{2\left(n-\frac{1}{2}\right)}{\eta\left(1-\eta\right)}\left(\xi-\eta\right)-\frac{n}{\eta\left(1-\eta\right)}\left(\xi-\eta\right)^2\right]\right\},$$

where we have set $\eta = H/N$ and $\xi = h/n$ (i.e. the percentage of white balls in the urn, and the frequency of white balls drawn in a sample of n, respectively). In the special case $\eta = \frac{1}{2}$ ($H = \frac{1}{2}N$; half the balls in the urn are white, half black), the expression simplifies considerably to give

$$\exp\left\{-\frac{1}{2}\frac{n}{N}\left[1-4n\left(\xi-\frac{1}{2}\right)^2\right]\right\}.$$

On the basis of this, we conclude that (approximately) the distribution gives higher probabilities than the Bernoulli distribution in the range where h lies between $\eta\eta \pm \sqrt{[n\eta(1-\eta)]}$ (i.e. between $m \pm \sigma$), with a maximum at $n\eta$, and lower values outside this interval.

It is possible, however, to avoid the rather tiresome details (which we have skipped over) by using the argument already encountered in Chapter 4, 4.17.5, and observing that

$$\sigma^2(S_n) = nq\tilde{q} + 2\binom{n}{2}rq\tilde{q} = nq\tilde{q}\left[1 + (n-1)r\right].$$

For $n = N$, we have $\sigma(S_N) = 0$, because $(S_N = H)$ = the certain event, and hence

$$1 + (N-1)r = 0, \qquad r = -1/(N-1). \tag{7.8}$$

Remarks. Note how useful it can be to bear in mind that apparently different problems may be identical, and how useful it can be to have derived different forms of expression for some given result, to have found their probabilistic interpretations, and to be in a position to recognize and use the simplest and most meaningful form in any given situation. Note, also, that one should be on the lookout for the possibility of reducing more complicated problems to simpler ones, both heuristically and, subsequently, by means of rigorous, detailed, analytical arguments, either exact or approximate.

In the case considered, we can go further and note that, by virtue of their interpretations within the problem itself, we have $\omega_h^{(n)} = \omega_{H-h}^{(N-n)}$. In other words, for given N and H, the distributions for complementary sample sizes n and $N - n$ are identical if we reverse $h = 0, 1, \ldots, H$ to $H, \ldots, 1, 0$ (and this can be seen immediately by glancing at the formula). It follows that, among other things, what is claimed to hold for 'small n must also hold for large n' (i.e. n close to N). The approximation does not work for central values ($n \sim \frac{1}{2}N$) and we note, in particular, that, for n lying between N and $N - H$, not all the values $h = 0, 1, \ldots, n$ are possible (since they themselves must lie between H and $n - (N - \mathrm{H})$.

7.4.4. *The Pascal distribution.* This is the distribution of X = 'the number of tosses required before the rth Head is obtained' (more generally, it arises for any independent events with arbitrary, constant probability p). Alternatively, it is the distribution of X such that $S_X = r > S_{X-1}$. By changing the scale, we could, of course, consider $X' = a + bX$. One example of this which often crops up is $X' = X - r$ = 'the number of failures preceding the rth success', but many of the other forms, such as those considered in the previous case, do not make sense in this context.

A new feature is that the distribution is unbounded, the possible values being $x_h = h = r, r + 1, r + 2, \ldots$ (up to infinity, and, indeed, $+\infty$ must be included as a possible value, along with all the integers, since it corresponds to the case where the infinite set of trials result in less than r successes). In line with our previous policy, we shall avoid critical questions by deciding that if the rth success is not obtained within a maximum of N trials (where N is very large compared with the other numbers in question) we shall set $X = N$. (To be precise, we shall consider $X' = X \wedge N$ instead of X) Were we to consider $X' = X - r$, the possible values would be $0, 1, 2, \ldots$ (and this is one of the reasons why this formulation is often preferred; another reason will be given in the discussion that follows; see equation 7.15).

For each r and p, we have, of course, a different distribution:

$$p_h = \binom{h-1}{r-1} p^r \tilde{p}^{h-r} \qquad \left(\text{if } p = \frac{1}{2}, p_h = \binom{h-1}{r-1}/2^h\right). \tag{7.9}$$

In fact (as we saw in (F), 7.2.4, for the special case $p = \frac{1}{2}$ in order to obtain $X = h$, we must have $r - 1$ successes in the first $h - 1$ trials, together with a success on the hth trial.

In terms of the random process representation, we are dealing with the crossing of the line $y = 2r - x$ (or, if one prefers, with the mass that ends up there if the line acts as an absorbing barrier).

Note that the series $\sum p_h$ sums up to 1. It must, of course, be convergent with sum $\leqslant 1$ by its very meaning; the fact that it is = 1, and not < 1, ensures that, as n increases, the probability that $X > n$ tends to zero (and, in particular, the probability that $X = \infty$ is 0).

So far as the behaviour of the distribution is concerned, the p_h again increase until they reach a maximum (attained for the greatest $h \leqslant r/p$), and then decrease to zero (asymptotically, like a geometric progression with ratio $\tilde{p}$). A more intuitive explanation of the increase is that it continues so long as the prevision of the number of successes, $\mathbf{P}(S_h) = hp$, does not exceed the required number of successes, r.

The geometric distribution. For $r = 1$, the Pascal distribution reduces to the special case of the geometric distribution:

$$p_h = p\tilde{p}^{h-r}, \tag{7.10}$$

forming a geometric progression ($p_1 = p$, with ratio $\tilde{p} = 1 - p$). If, for example, the first failure corresponds to elimination from a competition, this gives the probability of being eliminated at the hth trial, when the probability of failure at each trial is p. (N.B. For the purposes of this particular example, we have, for the time being, interchanged 'success' and 'failure'.) In particular, it gives the probability that a machine first goes wrong the hth time it is used, or that a radioactive atom disintegrates in h years time and so on, where the probability of occurrence is p on each separate occasion. (If the probability of death were assumed to be constant, rather than increasing with age, this would also apply to the death of an individual in h years time.)

The property of giving the same probability, irrespective of the passing of time, or of the outcomes of the phenomenon in the past, is known as the *lack of memory* property of the geometric distribution. The waiting time for a particular number to come up on the lottery[22] has, under the usual assumptions, a geometric distribution (the ratio is $\tilde{p} = \left(\frac{17}{18}\right) = 94 \cdot 44\%$, for a single city; $\tilde{p} = \left(\frac{17}{18}\right)^{10} = 56\%$ for the whole set of ten cities). This provides further confirmation, if such were needed, of the absurdity of believing that numbers which have not come up for a long time are more likely to be drawn in future.

To put this more precisely: it is absurd to use the small probabilities of long waiting times, which are themselves evaluated on the basis of the usual assumptions, and are given by the geometric distribution (or to invoke their comparative rarity, statistically determined in accordance with it), to argue, on the basis of independence, against the very assumptions with which one started – that is the *lack of memory* property. If, on the other hand, someone arrived at a coherent evaluation of the probabilities by a different route, we might not judge him to be reasonable, but this would simply be a matter of opinion.

Finally, let us give the explicit expression for the case $r = 2$ (again, this could be thought of as elimination from a competition, but this time at the second failure): it reduces to $p_h = (h-1)p^2\tilde{p}^{h-2}$.

---

22 *Translators' note.* See footnote 28 in Chapter 2.

Prevision and standard deviation. In order to calculate $\mathbf{P}(X)$ and $\boldsymbol{\sigma}(X)$, it turns out to be sufficient to do it for the case $r = 1$. We obtain

$$\mathbf{P}(X) = p\sum_{h=1}^{\infty} h\tilde{p}^{h-1} = p/(1-\tilde{p})^2 = 1/p, \tag{7.11}$$

$$\begin{aligned}\mathbf{P}(X^2) &= p\sum_{h=1}^{\infty} h^2\tilde{p}^{h-1} = p\tilde{p}\sum_{h=1}^{\infty} h(h-1)\tilde{p}^{h-2} + p\sum_{h=1}^{\infty} h\tilde{p}^{h-1} \\ &= 2p\tilde{p}/p^3 + 1/p = (2-p)/p^2\end{aligned} \tag{7.12}$$

(verify this!), and hence,

$$\sigma^2(X) = \mathbf{P}(X^2) - \mathbf{P}^2(X) = (1-p)/p^2.$$

For general r, it suffices to note that

$$\mathbf{P}(X) = r/p, \tag{7.13}$$

$$\sigma^2(X) = r(1-p)/p^2, \tag{7.14}$$

because (as we already observed for $\mathbf{P}(X)$ in the case $p = \frac{1}{2}$; see (*M*), 7.2.7) we can consider X as the sum of r terms, $X_1 + X_2 + \ldots + X_r$, stochastically independent, and each corresponding to $r = 1$ (X_i = the number of trials required after the $(i - 1)$th failure until the ith failure occurs).

Comments. This technique will also be useful in what follows: note that it can also be used for $X' = X - r$ if we consider r summands of the form $X_i' = X_i - 1$.

In this context (i.e. with h transformed into $h + r$), the p_h are given by

$$p_h = \binom{h+r-1}{r-1} p^r\tilde{p}^h = \binom{h+r-1}{h} p^r\tilde{p}^h = (-1)^h\binom{-r}{h} p^r\tilde{p}^h, \tag{7.15}$$

where the definition $\binom{x}{h} = x(x-1)\ldots(x-h+1)/h!$ is extended to cover any real x (not necessarily integer, not necessarily positive).

If we do this, the distribution then makes sense for any real $r > 0$. This generalized form of the Pascal distribution (which has integer r) is called the *negative binomial distribution* (simply because it involves the notation $\binom{-r}{h}$). For $r = 0$, the distribution is concentrated at the origin ($p_0 = 1$, $p_h = 0$, $h \neq 0$); for $r \sim 0$, we have $p_h \simeq r\tilde{p}^h/h$ ($h \neq 0$) (the logarithmic distribution; see Chapter 6, 6.11.2), and hence

$$p_0 \simeq 1 - r\sum_{h=1}^{\infty} \tilde{p}^h/h = 1 - r\log(1/p). \tag{7.16}$$

We shall make use of this later on, and the significance of the extension to noninteger r will also be explained.

The prevision in this case is clearly given by

$$\mathbf{P}(X') = \mathbf{P}(X) - r = r\tilde{p}/p, \tag{7.17}$$

whereas the standard deviation, $\sqrt{(r\tilde{p})}/p$, is unaltered; this also holds for noninteger r.

Another form. We have already seen (Section 7.3) that, if $S_N = H$ is assumed to be known, the problem of the location, h, of the rth success leads to the distribution

$$p_h = \binom{h-1}{r-1}\binom{N-h}{H-r}/\binom{N}{H}, \tag{7.18}$$

rather than to the Pascal distribution. For an example where this distribution occurs, consider an election in which N votes have been cast and are counted one at a time. Suppose further that a candidate is to be declared elected (or a thesis accepted) when r votes in favour have been counted. Given that a total of $H \geqslant r$ out of N were actually favourable, equation 7.18 gives the probability that success is assured by the counting of the hth vote. (Another example, with $N = 90$ and $H = r = 15$, is given by the probability of completing a 'full house' at bingo with the hth number called.)

We shall restrict ourselves to considering the particularly simple case of $H = r = 1$, a case which is nonetheless important (and a study of the general case is left as an exercise for the reader). Clearly (even without going through the algebra), we have $p_h = 1/N$ for $h = 1, 2,\ldots, N$. If there has only been one success in N trials (or there is only one favourable vote in the ballot box, or only one white ball in the urn, or only one ball marked '90'), there is exactly the same probability of finding it on the first, second,..., or Nth (final) trial.

7.4.5. *The discrete uniform distribution*. This is the name given to the distribution of an X which can only take on a finite number of equally spaced possible values, each with the same probability: for example $x_h = h$, $h = 1, 2,\ldots, n$ (or $x_h = a + bh$), with all the $p_h = 1/n$. As examples, we have a fair die ($n = 6$), a roulette wheel ($n = 37$) or the game of bingo ($n = 90$).

It is easily seen that

$$\mathbf{P}(X) = \frac{1}{2}(n+1), \mathbf{P}(X^2) = \left(1^2 + 2^2 + \ldots + n^2\right)/n = \left(4n^2 + 6n + 2\right)/12,$$

from which (subtracting $\left[\frac{1}{2}(n+1)\right]^2 = \left(3n^2 + 6n + 3\right)/12$) we obtain

$$\sigma^2(X) = \left(n^2 + 1\right)/12, \quad \sigma(X) = \left(n/\sqrt{12}\right)\sqrt{\left(1 + 1/n^2\right)} \simeq n/\sqrt{12}.$$

A random process (*Bayes–Laplace*, *Pólya*). Using this distribution as our starting point, we can develop a random process similar to that which led to the hypergeometric distribution. In fact, we consider successive drawings (without replacement) from an urn containing N balls, with the possible number of white balls being any of 0, 1, 2,..., N, each with probability $1/(N+1)$. (This could arise, for example, if the urn were chosen from a set of $N + 1$ urns, ranging over all possible compositions, and there were no grounds for attributing different probabilities to the different possible choices.)

Let us assume, therefore, that $\omega_H^{(N)} = 1/(N+1)(H = 0,1,2\ldots,N)$, and that (as in the case of a known composition, H/N) all the permutations of the possible orders of drawing the balls are equally probable: in other words, that all the dispositions of H white and $N - H$ black balls (i.e. all the paths from 0 ending up at the same final point $[N, H]$) are equally probable. Each of these paths therefore has probability $1/\binom{N}{H}(N+1)$.

We shall now show that, under these conditions, the distribution for every S_n ($n < N$) is uniform, just as we assumed it to be for S_N: that is

$$\omega_h^{(n)} = 1/(n+1) \quad (h = 0, 1, 2, \ldots, n).$$

It can be verified, in a straightforward but tedious fashion, that

$$\sum_{H=0}^{N} \binom{n}{h}\binom{N-n}{H-n} \frac{1}{\binom{N}{H}(N+1)} = \frac{1}{n+1}$$

(the probabilities of the paths terminating at $[N, H]$ multiplied by the number of them that pass through $[n, h]$, the sum being taken over H). The proof by induction (from N to $N - 1$, $N - 2$,..., etc.) is much simpler, however, and more instructive. It will be sufficient to establish the step from N to $N - 1$. The probability $\omega_h^{(N-1)}$ that $S_{N-1} = h$ is obtained by observing that this can only take place if $H = h$ and the final ball is black, or if $H = h + 1$ and the final ball is white; each of these two hypotheses has probability $1/(N + 1)$ and the probabilities of a black ball under the first hypothesis and a white ball under the second are given by $(N - h)/N$ and $(h + 1)/N$, respectively. It follows that

$$\omega_h^{(N-1)} = \frac{1}{N+1}\left(\frac{N-h}{N} + \frac{h+1}{N}\right) = \frac{1}{N}. \tag{7.19}$$

Expressed in words: if all compositions are equally probable, so are all the frequencies at any intermediate stage. This property ($\omega_h^{(n)} = 1/(n+1)$) can also hold for all n (without them being bounded above by some pre-assigned N), and leads to the important Bayes–Laplace process (which we shall meet in Chapter 11, 11.4.3) or, with a different interpretation, to the Pólya process (Chapter 11, 11.4.4) with which it will be compared.

7.5 Laws of 'Large Numbers'

7.5.1. We now return to our study of the random process of Heads and Tails (as well as some rather less special cases) in order to carry out a preliminary investigation of what happens when we have 'a large number' of trials. This preliminary investigation will confine itself to qualitative aspects of the order of magnitude of the deviations. In a certain sense, it reduces to simple but important corollaries of an earlier result, which showed that *the order of magnitude increases as the square root of n* (the number of trials).

In the case of Heads and Tails $\left(p = \frac{1}{2}\right)$ the prevision of the *gain*, Y_n, is zero (the process is fair; $\mathbf{P}(Y_n) = 0$), and its standard deviation $\boldsymbol{\sigma}(Y_n)$ (which, in a certain sense, measures 'the order of magnitude' of $|Y_n|$) is equal to $\sqrt{n}$. The *number of successes* (Heads) is denoted by S_n and has prevision $\frac{1}{2}n$; its standard deviation (the order of magnitude, measured by σ) is equal to $\frac{1}{2}\sqrt{n}$. For the *frequency of successes*, S_n/n, the prevision and standard deviation are those we have just given, but now divided by n; that is $\frac{1}{2}$ and $\frac{1}{2}/\sqrt{n}$, respectively. In a similar way, one might be interested in the *average gain* (per toss), Y_n/n; this has prevision 0 and standard deviation $1/\sqrt{n}$.

The fact that

$$\mathbf{P}\left[\left(\frac{S_n}{n}-\frac{1}{2}\right)^2\right]=\frac{1}{4n}\to 0$$

is expressed by saying that *the frequency converges in mean-square to* $\frac{1}{2}$. This implies (see Chapter 6, 6.8.3) that it also converges *in probability*. Similarly, the average gain converges to 0 (both in mean-square and in probability).

We recall that convergence in probability means that, for positive ε and θ (however small), we have, for all n greater than some N,

$$\mathbf{P}\left(\left|\frac{S_n}{n}-\frac{1}{2}\right|>\varepsilon\right)<\theta.$$

A straightforward application of Tchebychev's inequality shows that the probability in our case is less than $\sigma^2/\varepsilon^2=1/4n\varepsilon^2$.

7.5.2. When referring to this result, one usually says, in an informal manner, that after a large number of trials it is *practically certain* that the frequency becomes *practically equal* to the probability. Alternatively, one might say that 'the fluctuations tend to cancel one another out'. One should be careful, however, to avoid exaggerated and manifestly absurd interpretations of this result (a common trap for the unwary). Do not imagine, for example, that convergence to the probability is to be expected because future discrepancies should occur in such a way as to 'compensate' for present discrepancies by being in the opposite direction. Nor should one imagine that this holds for the absolute deviations. It is less risky to gamble just a few times (e.g. ten plays at Heads and Tails at 1000 lire a time) than it is to repeat the same bet many times (e.g. a 1000 plays are 10 times more risky; $10 = \sqrt{(1000/10)}$). On the other hand, it would be less risky to bet 1000 times at 10 lire a time. Furthermore, if at a certain stage one is losing – let us say 7200 lire – the law of large numbers provides no grounds for supposing that one will 'get one's own back'.[23] In terms of prevision, this loss remains forever at the same level, 7200 lire. The future gain (positive or negative) has prevision zero but, as one proceeds, the order of magnitude becomes larger and larger and, eventually, it makes the loss already suffered *appear negligible*. It is in this sense, and only in this sense, that the word 'compensate' might reasonably be used, since one would then avoid the misleading impression that it usually conveys. The fact remains, however, that the loss has already been incurred.

Observe once again how absurd it would be to imagine, a priori, some sort of correlation – which would be a consequence of laws and results derived on the basis of an assumption of independence!

7.5.3. In addition to this, one should note that the property we have established concerns the probability of a deviation $>\varepsilon$ between the probability and the frequency for

23 The illusory nature and pernicious influence of such assumptions are referred to in a popular, witty saying (possibly Sicilian in origin), in itself rather remarkable, given that popular prejudice seems on the whole to incline towards the opposite point of view. The saying concerns the answer given by a woman to a friend, who has asked whether it was true that her son had lost a large amount of money gambling: '*Yes, it's true*', she replies, '*But that's nothing: what is worse is that he wants to get his own back*!'.

an individual n (although it can be for any $n \geqslant N$). This clearly does not imply – although the fact that it does not can easily escape one's notice – that the probability of an 'exceptional' deviation occurring at least once for an n between some N and an $N + K$ greater than N is also small.

It is especially easy to overlook this if one gets into the habit of referring to events with small probability as 'impossible' (and even worse if one appears to legitimize this bad habit by giving it a name – like 'Cournot's principle', Chapter 5, 5.10.9). In fact, if an 'exception' were impossible for every individual case $n \geqslant N$, it would certainly be impossible to have even a single exception among the infinite number of cases from $n = N$ onwards.

If one wanted to use the word 'impossible' in this context without running into these problems, it would be necessary to spell out the fact that it should not be understood as meaning 'impossible', but rather 'very improbable'. However, anyone who states that 'horses are potatoes', making it clear that when it refers to horses the meaning of 'potato' is not really that of 'potato' (but rather that of 'horse'), would probably do better not to create useless terminological complications in the first place (since, in order for it not to be misleading, it must be immediately followed by a qualification which takes away its meaning).

Now let us return to the topic of convergence. If the probability of a deviation $|S_n/n - \frac{1}{2}|$ at Heads and Tails being greater than ε were actually equal to $1/4n\varepsilon^2$, then, for any N, in some interval from N to a sufficiently large $N + K$ the prevision of the number of 'exceptions' (deviations $>\varepsilon$) would be arbitrarily large (approximately $(1/4\varepsilon^2)\log(1 + K/N)$). This follows from the fact that the series $\sum 1/n$ is divergent and the sum between N and $N + K$ is approximately equal to $\log(1 + K/N)$. In fact, as we shall see later, the result we referred to at the beginning of 7.5.3 does hold. It just so happens that the Tchebychev inequality, although very powerful in relation to its simplicity, is not sufficient for this more delicate result. Stated mathematically, we have, for arbitrary positive ε and θ,

$$\mathbf{P}\left(\max_{N \leqslant n \leqslant N+K} \left| \frac{S_n}{n} - \frac{1}{2} \right| > \varepsilon \right) < \theta,$$

provided N is sufficiently large (K is arbitrary).[24] This form of stochastic convergence is referred to as *strong convergence* and the result is known as the '*strong law of large numbers*'. By way of contrast, the word 'strong' is replaced by '*weak*' when we are referring to convergence in probability, or to the previous form of the law of large numbers.

7.5.4. In order to fix ideas, we have referred throughout to the case of Heads and Tails. Of course, the results also hold for $p \neq \frac{1}{2}$ (except that we then have to write $\sigma^2 = p\tilde{p}$, which is $< \frac{1}{4}$ unless $p = \frac{1}{2}$) and even in the case where the $p_i = \mathbf{P}(E_i)$ vary from event to event (provided $\sum p_i \tilde{p}_i$ diverges, which may not happen if the p_i get too close to the extreme values 0 and 1). In the latter case, our statement would, in general, assert that the difference between the frequency S_n/n in the first n trials and the arithmetic mean of the probabilities, $(p_1 + p_2 + \ldots + p_n)/n$, tends to zero (in mean-square and in probability). Only if the arithmetic mean tends to a limit p (or, as analysts would say, if the p_i are a

24 Were it not for our finitistic scruples (see Chapter 6 and elsewhere), we could do as most people do, and write $\sup(n \geqslant N)$ in place of $\max(N \leqslant n \leqslant N + K)$, saying that it is 'almost certain' (i.e. the probability = 1) that $\lim(S_n/n) = \frac{1}{2}$ (in the sense given).

sequence converging to *p in the Cesàro sense*) does the previous statement in terms of deviation from a fixed value hold (and the fixed value would then be the limit *p*).

We can, however, say a great deal more on the basis of what we have established so far. The only properties we have made use of are those of the previsions and standard deviations of the gains of individual bets, $2E_i - 1$, and of their sums, Y_n. It is easy to convince oneself that for the conclusion (weak convergence) to hold we only require that the gains X_i ($i = 1, 2,\ldots$) have certain properties. For example, it suffices that they have zero prevision and are (*pairwise*) *independent with constant, finite standard deviations*. More generally, we only require that they are (pairwise) uncorrected and that the standard deviations $\boldsymbol{\sigma}(X_i) = \sigma_i$ are bounded, and such that $\sum\sigma_i^2$ diverges. Considering the case of zero prevision for convenience, we have

$$Y_n/n = \left(X_1 + X_2 + \ldots + X_n\right)/n \dot{\to} 0 \quad \left(\text{and hence} \overset{<}{\to} 0\right).$$

Expressed in words: the (weak) law of large numbers holds for sums of uncorrelated random quantities under very general conditions, in the sense that *the arithmetic mean*, Y_n/n, *tends to* 0 *in quadratic prevision, and the probability of its having an absolute value* $>\varepsilon$ (*an arbitrary, preassigned positive value*) *also tends to* 0.

The *strong law of large numbers* also holds under very general conditions. The argument which ensures its validity if the sum of the probabilities p_h of 'exceptions' (deviations $|Y_h/h| > \varepsilon$) converges, turns out to be sufficient if these probabilities are evaluated on the basis of the normal distribution, and this will be the case if the X_h are assumed to be independent with standard unit normal distributions ($m = 0$, $\sigma = 1$). Asymptotically, however, this property also holds in the case of Heads and Tails, and for any other X_h which are identically distributed with finite variances (let us assume $\sigma = 1$).[25] We shall, as we mentioned above, restrict ourselves to the proof based on the convergence of Σp_h. Afterwards, we shall mention the possibility of modifications which make the procedure much more powerful.

Since the distribution function of the (standard unit) normal cannot be expressed in a closed form, it is necessary, in problems of this kind, to have recourse to an asymptotic formula (which can easily be verified – by L'Hospital's rule, for example):

$$1 - F(x) = \frac{1}{\sqrt{(2\pi)}}\int_x^{\infty} \mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \sim \frac{1}{\sqrt{(2\pi)}x}\mathrm{e}^{-\frac{1}{2}x^2} \quad \left(\text{as } x \to +\infty\right). \tag{7.20}$$

It follows, since Y_h has standard deviation $\sqrt{h}$, that $|Y_h/h| > \varepsilon$ can be thought of as $|Y_h/\sqrt{h}| > \varepsilon\sqrt{h} = \varepsilon\sqrt{h} \times$ the standard deviation of the standardized distribution, and, therefore,

$$\begin{aligned} p_h &= 1 - F\left(\varepsilon\sqrt{h}\right) - F\left(-\varepsilon\sqrt{h}\right) = 2\left[1 - F\left(\varepsilon\sqrt{h}\right)\right] \\ &\sim \frac{2}{\sqrt{(2\pi)}\varepsilon\sqrt{h}}\mathrm{e}^{-\frac{1}{2}h\varepsilon^2} = \frac{K}{\sqrt{h}}\mathrm{e}^{-ch}. \end{aligned}$$

25 That the normal distribution frequently turns up is a fact which is well known, even to the layman (though the explanation is often not properly understood). The case we are referring to here will be dealt with in Section 7.6; we shall not, therefore, enter into any detailed discussion at present.

The geometric series $\sum e^{-ch}$ converges, however, and hence, *a fortiori*, so does the series $\sum p_h$, with the terms divided by $\sqrt{h}$. The remainder, from some appropriate N onwards, is less than some preassigned θ, and this implies that the probability of even a single 'exceptional deviation', $|Y_h/h| > \varepsilon$, for h lying between N and an arbitrary $N + M$, is less than θ. (If countable additivity were admitted, one would simply state the result 'for all $h \geqslant N$'.)

The conclusion can easily be strengthened by observing that convergence still holds even if the constant ε is replaced by some $\varepsilon(h)$ decreasing with h; for example,

$$\varepsilon(h) = \sqrt{\left(2a \ \log h\right)} / \sqrt{h}, \quad \text{with} \quad a > 1.$$

We then have $h\varepsilon^2 = 2a \log h$, and

$$p_h = \mathbf{P}\left(\left|Y_h / h\varepsilon\right| > (h)\right) = \left[K / \sqrt{\left(2a \ \log \ h\right)}\right] \ e^{-a \log h} = (\ldots) h^{-a}.$$

But the terms (...) tend to zero, the series $\sum h^{-a}$ ($a > 1$) converges, and, *a fortiori*, Σp_h converges. Expressed informally, this implies that, from some N on, it is 'almost certain that Y_h will remain between $\pm c\sqrt{(2h \log h)}$' for $c > 1$.

The argument that follows exemplifies the methods that could be used to further strengthen the conclusions. Indeed, we shall see precisely how it is that one arrives at a conclusion which is, in a certain sense, the best possible ('we shall see', in the sense that we will sketch an outline of the proof without giving the details).

We note that were we to consider only the possible exceptions (Y_h lying outside the interval given above) at the points $h = 2^k$, instead of at each h, we could obtain the same convergent series by taking

$$\varepsilon(h)\sqrt{h} = \sqrt{\left(2a \ \log \ k\right)} \sim \sqrt{\left(2a \ \log \ \log \ h\right)}, \quad \text{instead of} \sqrt{\left(2a \ \log \ h\right)}.$$

A conclusion that only applies to the values $h = 2^k$ is, of course, of little interest, but it is intuitively obvious that we certainly do not require a check on all the h. The graph of $y = Y(h)$ can scarcely go beyond the preassigned bounds if one checks that it has remained within them by scanning a sufficiently dense sequence of 'check points'. Well, one can show that the check points $h = 2^k$ (for example) are sufficiently dense for one to be able to conclude that – again expressed informally – it is almost certain that all the Y_h from some N on (an N which cannot be made precise), will, in fact, remain within much smaller bounds of the form $\pm c\sqrt{(2h \log \log h)}$, for $c > 1$.

What makes this result important is that, conversely, if $c < 1$, it is 'practically certain that these bounds will be exceeded, however far one continues'. This is the celebrated *law of the iterated logarithm*, due to Khintchin.

Note that, in order to prove the converse which we have just stated, the divergence of the series is not sufficient, unless the events are independent (Borel–Cantelli lemma). In the case under consideration, we do not have independence. We do, however, have the possibility of reducing ourselves to the latter case, because, if h'' is much larger than h' the contribution of the increment between h' and h'' (which is independent of $Y(h')$) is the dominating term in $Y(h'') = Y(h') + [Y(h'') - Y(h')]$.

All these problems can be viewed in a more intuitive light (and can be dealt with using other techniques, developed on the basis of other approaches) if we base ourselves on

random processes on the real line (and, so far as the results we have just mentioned are concerned, in particular on the Wiener–Lévy process). It will be a question of studying the graph $y = Y(t)$ of a random function in relation to regions like $|y| \leqslant y(t)$ (a preassigned function), by studying the probability of the graph entering or leaving the region, either once, or several times, or indefinitely.

Finally, let us just mention the standard set of conditions which are sufficient for the validity of the strong law. The X_h are required to be independent, and such that $\sum \sigma_h^2 / h^2$ converges (the Kolmogorov condition). The proof, which is based on an inequality due to Kolmogorov, one which, in a certain sense, strengthens the Tchebychev inequality and on the truncation of the 'large values' of the X_h, goes beyond what we wanted to mention at this stage.

In the classical case (that of independent events with equal probabilities), the weak and strong laws of large numbers are also known as the Bernoulli and Cantelli laws, respectively.

7.5.5. *The meaning and value of such 'laws'*. In addition to their intrinsic meaning, both mathematically and probabilistically, the laws of large numbers, and other asymptotic results of this kind, are often assigned fundamental rôles in relation to questions concerning the foundations of statistics and the calculus of probability. It seems appropriate to provide some discussion of this fact, both in order to clarify the various positions, and, in particular, to clarify our own attitude.

For those who seek to connect the notion of probability with that of frequency, results which relate probability and frequency in some way (and especially those results like the 'laws of large numbers') play a pivotal rôle, providing support for the approach and for the identification of the concepts. Logically speaking, however, one cannot escape from the dilemma posed by the fact that the same thing cannot both be assumed first as a definition and then proved as a theorem; nor can one avoid the contradiction that arises from a definition which would assume as certain something that the theorem only states to be very probable. In general, this point is accepted, even by those who support a statistical-frequency concept of probability; the attempts to get around it usually take the form of singling out, separating off, and generally complicating, particular concepts and models.

An example of this is provided by the 'empirical law of chance'. A phrase created for the purpose of affirming the *actual occurrence* of something the 'law of large numbers' states to be *very probable* comes to be presented as an *experimental fact*. Another example is provided by 'Cournot's principle': this states, as we mentioned in Chapter 5, 5.10.9, that 'an event of small probability does not occur', and covers the above, implicitly, as a special case. Sometimes, the qualification 'never, or almost never' is added, but although this removes the absurdity, in doing so it also takes away any value that the original statement may have had.

In any case, this kind of thing does nothing to break the vicious circle. It only succeeds in moving it somewhere else, or disguising it, or hiding it. A veritable labour of Sisyphus! It always ends up as a struggle against *irresolvable* difficulties, which, in a well-chosen phrase of B.O. Koopman, '*always retreat but are never finally defeated*, unlike Napolean's Guard'.

In order for the results concerning frequencies to make sense, it is necessary that the concept of probability, and the concepts deriving from it that appear in the statements and proofs of these results, should have been defined and given a meaning beforehand.

In particular, a result that depends on certain events being uncorrelated, or having equal probabilities, does not make sense unless one has defined in advance what one means by the probabilities of the individual events. This requires that probabilities are attributed to each of the given events (or 'trials'), that these all turn out to be equal and that, in addition, probabilities are attributed to the products of pairs of events, such that these are all equal and, moreover, equal to the square of the individual probabilities. In using the word 'attributed', we have, of course, used a word which fits in well with the subjectivistic point of view; in this context, however, it would make no difference if we were to think of such probabilities as 'existing', in accordance with the 'logical' or 'necessary' conception. In fact, the criticisms of the frequentistic interpretation made by Jeffreys, for instance, and the case against it which he puts forward (closely argued, and, I would say, unanswerable[26]), are in complete accord with the views we have outlined above. We acknowledge, of course, that there are differences between the necessary and subjectivistic positions (the latter denies that there are logical grounds for picking out *one single* evaluation of probability as being *objectively* special and 'correct'), but we regard this as of secondary importance in comparison with the differences that exist between, on the one hand, conceptions in which probability is probability (and frequency is just one of the ingredients of the 'outside world' which might or might not influence the evaluation of a probability) and, on the other hand, conceptions in which probability is, to a greater or lesser extent, a derivative of frequency, or is an idealization or imitation of it.

7.5.6. From our point of view, the law of large numbers forms yet another link in the chain of properties which justify our making use of expected or observed frequencies in our (necessarily subjective) evaluations of probability. We now see how to make use of the prevision of a frequency in this connection. The law of large numbers says that, *under certain conditions*, the value of the probability is *not only equal to the prevision* $\mathbf{P}(X)$ of a frequency X, but, moreover, *we are almost certain that X will be very close to this value* (getting ever closer, in a way that can be made precise, as one thinks of an even larger number of events).

This really completes the picture for the special case we have considered. Rather than introducing new elements into the situation (something we shall come across when we deal with exchangeable events in Chapter 11, and in similar contexts), we shall use these results in order to consider rather more carefully the nature of this special case: that is independent events with equal probabilities. It is important to realize that these assumptions, so apparently innocuous and easily accepted, contain unsuspected implications. To judge a coin to be 'perfectly fair so far as a single toss is concerned', means that one considers the two sides to be equally probable on this (the first) toss, or on any other toss for which one does not know the outcomes of the previous tosses. To judge a coin to be 'perfectly fair so far as the random process of Heads and Tails is concerned' is a very different matter.

The latter is, in fact, an extremely rash judgement that commits one to a great deal. It *commits* one, for example, to evaluating the probabilities as $\frac{1}{2}$ at each toss, *even if all*

26 See. Harold Jeffreys, *Scientific Inference*, Cambridge University Press; 1st edn (1931), 2nd edn (1957) and *Theory of Probability*, Oxford University Press; 1st edn (1938), 2nd edn (1948), 3rd edn (1961). Particularly relevant are the following: Section 9.21 of the first work, entitled 'The frequency theories of probability', and Sections 7.03–7.05 of the second, in Chapter 7, 'Frequency definitions and direct methods'.

the previous tosses (a thousand, or a million, or 10^{1000},...) *were all Heads*, or Heads and Tails alternately, and so on. Another consequence (although one which is well beyond the range of intuition) is that, for a sufficiently large number of tosses, one considers it advantageous to bet on the frequency lying between 0·49999 and 0·50001 rather than elsewhere in the interval [0, 1] (and the same holds for $0{\cdot}5 \pm 10^{-1000}$ etc.). I once remarked that '*the main practical application of the law of large numbers consists of persuading people how unrealistic and unreasonable it is, in practice, to make rigid assumptions of stochastic independence and equal probabilities*'. The remark was taken up by L.J. Savage, to whom it was made, and given publicity in one of his papers. It was intended to be witty, in part facetious and paradoxical, but I think that it is basically an accurate observation.

Notwithstanding its great mathematical interest, there is clearly even less to be said from a realistic and meaningful point of view concerning the strong law of large numbers.

7.5.7. *Explanations based on 'homogeneity'.* First of all, it is necessary to draw attention to the upside-down nature of the very definitions of notions (or would-be notions) like those of *homogeneous* events, *perfect* coins and so on. Any definition that is framed in objective, physical terms, or whatever, is not suitable, because it cannot be used to *prove* that a given probabilistic opinion is a logical truth, nor can it *justify its imposition* as an article of faith.

If one wants to make use of these, or similar, notions, it is clear, therefore, that their meanings can only come about and be made precise as expressions of particular instances of probabilistic opinions (opinions which, had one already attributed to these notions a metaphysical meaning, preceding these personal opinions, one would have called a *consequence* of it).

I recall a remark, dating from about the time of my graduation, which has remained engraved upon my memory, having struck me at the time as being very accurate. A friend of mine used to say, half-jokingly, and in a friendly, mocking way, that it was never enough for me to define a concept, but that I needed to 'definettine' it. In actual fact, I had, by and large, adopted the mode of thinking advocated by authors like Vailati and Calderoni (or perhaps it would be more accurate to say that I found their approach to be close to my own). Papini used to say of Calderoni that 'what he wanted to do was to show what precautions one ought to take, and what procedures one ought to use, in order to arrive at statements which make sense'.[27] On the other hand, it was precisely this form of reasoning which, in successive waves, from Galileo to Einstein, from Heisenberg to Born, freed physics – and with it the whole of science and human thought – from those superstructures of absurd metaphysical dross which had condemned it to an endless round of quibbling about pretentious vacuities.

At the same time, and for the reasons we have just given, any attempt to define a coin as 'perfect' on the basis of there being no objective characteristics that prevent the probability of Heads from being $p = \frac{1}{2}$, or different tosses from being stochastically independent, is simply a rather tortuous way of making it appear that the above-mentioned objective circumstances play a decisive rôle. In fact, they are mere window dressing. The real meaning only becomes clear when these circumstances are pushed on one side and one simply proceeds as follows (and, in doing so, discovers that there are two possible meanings of 'perfect'): we shall use the expression *perfect coin* in the *weak* sense as a shorthand statement of the fact that we attribute equal probabilities $\left(\frac{1}{2}\right)$ to each of the

27 G. Papini, *Stroncature*, No. 14: 'Mario Calderoni'; G. Vailati, *Scritti* (in particular, see those works quoted in the footnotes to Chapter 11, 11.1.5).

two possible outcomes of a toss; we use the expression in the *strong* sense if we attribute equal probabilities $\left(\left(\frac{1}{2}\right)^n\right)$ to each of the 2^n possible outcomes of n tosses, for any n. This does not mean, of course, that in making such a judgment it is not appropriate (or, even less, that it is not admissible) to take into account all those objective circumstances that one considers relevant to the evaluation of probability. It merely implies that the evaluation (or, equivalently, the identification and listing of the circumstances that might 'reasonably' influence it) is not a matter for the theory itself, but for the individual applying it. From his knowledge of the theory, the individual will have at his disposal various auxiliary devices to aid him in sharpening his subjective analysis of individual cases; the standard schemes will serve as reference points for his idealized schemes. There is no way, however, in which the individual can avoid the burden of responsibility for his own evaluations. The key cannot be found that will unlock the enchanted garden wherein, among the fairy rings and the shrubs of magic wands, beneath the trees laden with monads and noumena, blossom forth the flowers of *Probabilitas realis*. With these fabulous blooms safely in our button-holes we would be spared the necessity of forming opinions, and the heavy loads we bear upon our necks would be rendered superfluous once and for all.

7.5.8. Having dealt with the logical aspect, it remains to consider, in a more detailed fashion, the criticisms of those discussions based upon *homogeneity*, both from a practical point of view and from the point of view of the 'realism' of such a notion in relation to actual applications. It is curious to observe that these kinds of properties (independence with equal probabilities) are even less realistic than usual in precisely those cases that correspond to the very empirical–statistical interpretation which claims to be the most 'realistic' (i.e. those attributing the 'stability of the frequency' to quasi-'physical' peculiarities of some phenomenon possessing 'statistical regularity').

Can we really believe that a coin – 'perfect' so far as we can see – provides the perfect example of a phenomenon possessing these 'virtues'? There appears to be room for doubt. Is it not indeed likely that 'suspicious' outcomes would lead us to re-evaluate the probability, somehow doubting its perfection, or the manner of tossing, or something else?

By way of contrast, we would have less reason for such suspicions and doubts if, from time to time, or even at each toss, the coin were changed. This would be even more true if coins of different denomination were used and the person doing the tossing were replaced, and more so again if the successive events considered were completely different in kind (for example: whether we get an even or odd number with a die, or with two dice, or in drawing a number at bingo, or for the number plate of the first car passing by, or in the decimal place of the maximum temperature recorded today and so on; whether or not the second number if greater than the first when we consider number plates of cars passing by, ages of passers-by, telephone numbers of those who call us up and so on; it is open to anyone to display their imagination by inventing other examples). Under these circumstances, it seems very unlikely that a 'suspicious' outcome, whatever it was, would lead one to expect similar strange behaviour from future events, which lack any similarity or connection with those that have gone before.[28]

28 We have used the qualification 'suspicious' only after careful consideration (we have avoided, for example, 'exceptional', or 'strange', or 'unlikely'). Now is not the appropriate time for a detailed examination of this question, however. This will come later (in Chapter 11, 11.3.1), and will clarify the meaning of this term and the reasons why it was chosen.

This demonstrates that the homogeneity of the events (the fact of their being, in some sense, 'trials of the same phenomenon', endowed with would-be statistical virtues of a special kind) is by no means a necessary prerequisite for the possible acceptance of the properties of independence with equal probabilities. In fact, it is a positive obstacle to such an acceptance. If, in such a case, the above properties are accepted, it is not that they should be thought of as valid *because of homogeneity* but, if at all, *in spite of homogeneity* (and it is much easier to accept them in other cases *because of heterogeneity*).[29] Despite all this, we continue to hear the exact opposite, repeated over and over, with the tiresome insistence of a silly catchphrase.

The 'laws of chance' (although it is rather misleading to refer to them in this way) express, instead, precisely that which one can expect from a maximum of disorder, in which any kind of useful knowledge is lacking. Any increase in one's knowledge of the phenomena and of their 'properties' would, if it were to be at all useful, lead one to favour some subset of the 2^n possible outcomes, and hence would lead to an evaluation of probabilities which are *better in this specific case* (with respect to the judgment of the individual who makes his evaluation after taking it into account) than those which would be valid in the absence of any information of this kind. There exists no information, knowledge, or property, that can strengthen or give 'physical' (or philosophical, or any other) meaning to the situation which corresponds to a perfect symmetry of ignorance.[30]

7.6 The 'Central Limit Theorem'; The Normal Distribution

7.6.1. If one draws the histograms of the distribution of Heads and Tails (the binomial distribution with $p = \frac{1}{2}$) and compares them for various values of n (the number of tosses), one sees that the shape remains the same (apart from discontinuities and truncation of the tails, features which arise because of the discreteness, and tend to vanish as n increases). The shape, in fact, suggests that one is dealing with the familiar normal distribution (the Gaussian distribution, or 'distribution of errors', which we mentioned briefly in Chapter 6, 6.11.3, and will further treat in Section 7.6.6; see Figure 7.6). Figure 7.5 gives the histogram for $n = 9$ (which is, in fact, a very small n!), together with the density curve. The agreement is already quite good, and the curve and the boundary of the histogram would rapidly become indistinguishable if we took a larger n (not necessarily very large).

In order to adjust the histograms to arrive at a unique curve, it is, of course, necessary to make an appropriate change of scale (we are concerned with convergence to a *type* of distribution; see Chapter 6, 6.7.1). The standard procedure of transforming to $m = 0$ and $\sigma = 1$ (Chapter 6, 6.6.6) is convenient and this is what we have done in the figure.[31]

---

29 A fuller account of this may be found in 'Sulla "compensazione" tra rischi eterogenei', *Giorn. Ist. Ital. Attuari* (1954), 1–21.

30 The following point has been made many times, and should be unnecessary. We are not speaking of exterior symmetries (which could exist), nor of 'perfect ignorance' (which cannot exist – otherwise, we would not even know what we were talking about), but about symmetry of judgment as made by the individual (in relation to the notion of indifference which he had prior to obtaining information, whether a great deal or only a small amount).

31 This holds for the actual distribution (discrete: the mass of every small rectangle concentrated at the centre). If one thinks of it as diffused, one must modify this slightly (an increase) as we shall see shortly (see 7.6.2, the case of F_n^I).

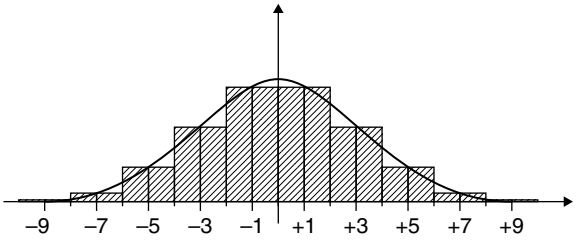


Figure 7.5 The binomial distribution: Heads and Tails $\left(p = \frac{1}{2}\right)$ with n tosses, $n = 9$. The possible values for the gain run through the ten odd numbers from –9 to +9, and the height of the column indicates the probability concentrated on each of these numbers. To give a more expressive picture, the values are assigned uniformly within ±1 of each point; this makes much clearer the approach of the binomial to the normal distribution, which will, in fact, be shown to be the limit distribution (as $n \to \infty$).

If we were, in fact, to consider the representation in terms of the natural scale (the gain Y_n, or the number of successes S_n) it would flatten out more and more, since the deviation behaves like $\sqrt{n}$. (One could think in terms of Bittering's apparatus, in which less and less sand would remain in each section as one continued to overturn it; see Figure 7.4: on the other hand, a large number of sections would be needed if one were to continue for very long.) In contrast to this, if one were to represent it in relative terms (the mean gain per toss, Y_n/n, or the frequency, S_n/n) the curve would shrink like $1/\sqrt{n}$, and would rise up to a peak in the centre. The remainder outside this central interval would tend to zero by virtue of the Tchebychev inequality. An appropriate choice is somewhere in between; as we have seen, one can take $Y_n/\sqrt{n}$ (i.e. the standardized deviation, both of the gain, and of the frequency from its prevision, $\frac{1}{2}$).

A somewhat more detailed study of the distribution of Heads and Tails will show us straightaway that the convergence to the normal distribution which is suggested by a visual inspection does, in fact, take place. In this case, too, however, the conclusions are valid more generally. They are valid not only for any binomial process with $p \neq 0$ (the effect of any asymmetry tends to vanish as n increases[32]) but also for sums of arbitrary, independent random quantities, provided that certain conditions (which will be given at the end of the chapter) are satisfied.

7.6.2. The limit distribution F of a sequence of distributions is to be understood in the sense defined in Chapter 6, 6.7.1: $F_n \to F$ means that, in terms of the distribution functions, $F_n(x) \to F(x)$ at all but at most a countable set of points (more precisely, at all but the possible discontinuity points of $F(x)$). This does not imply, of course, that if the densities exist we must necessarily also have $f_n(x) \to f(x)$; even less does it imply that if the densities are themselves differentiable we must have $f_n'(x) \to f'(x)$. Conversely, however, it is true that these properties do imply the convergence of the distributions (indeed, in a stronger and stronger, and intuitively meaningful way; one only has to think in terms of the graph of the density function).

32 We mention this case explicitly since many people seem to doubt it (notwithstanding the fact that it is clearly covered by the general theorem). Perhaps this is the result of a misleading prejudice deriving from too much initial emphasis on Heads and Tails (?).

In our case, we can facilitate the argument by first reducing ourselves to the case just mentioned (albeit with a little trickery along the way). In fact, our distributions are discrete, standardized binomial with $p = \frac{1}{2}$ and hence with probabilities $p_h = \binom{n}{h} / 2^n$ concentrated at the points $x_n = (2h - n)/\sqrt{n}$ (distance $2/\sqrt{n}$ apart, lying between $\pm\sqrt{n}$). In order to obtain a distribution admitting a density, it is necessary to distribute each mass, p_h, uniformly over the interval $x_h \pm 1/\sqrt{n}$; or, alternatively, with a triangular distribution on $x_h \pm 2/\sqrt{n}$. In this way, we obtain a continuous density (in the first case, the density is a step function; in the second, the derivative is a step function): we shall denote these two distributions by F_n^I and F_n^{II}, respectively.

We can also give a direct interpretation in terms of random quantities. The distributions arise if, instead of considering $Y_n/\sqrt{n}$, we consider $(Y_n + X)/\sqrt{n}$, where X is a random quantity independent of Y_n, having the appropriate distribution (in the cases we mentioned, $f(x) = \frac{1}{2}(|x| \leqslant 1)$, or $f(x) = \frac{1}{4}(2 - |x|)$ $(|x| \leqslant 2)$, respectively[33]). We observe immediately that, since no mass is shifted by more than $1/\sqrt{n}$ (or $2/\sqrt{n}$, respectively) in one direction or the other, F_n^I and F_n^{II} will, for each x, from some $n = N$ on, lie entirely between $F_n(x - \varepsilon)$ and $F_n(x + \varepsilon)$ (in fact, it suffices that $\varepsilon > 2/\sqrt{n}$; i.e. $n > N = 4/\varepsilon^2$). It follows that, so far as the passage to the limit is concerned, it will make no difference if we use these variants in place of the actual F_n (for notational convenience, we shall not make any distinctions in what follows; we shall simply write F_n. The change in the standard deviation also makes no difference, and can be obtained immediately, without calculation, from the previous representation: X has standard deviation $1/\sqrt{3}$ in the case of a uniform distribution (between ± 1), and $\sqrt{(2/3)}$ for a triangular distribution (between ± 2), and it therefore follows that the addition of X either changes the standard deviation to $\sqrt{(1 + 1/3n)}$ or to $\sqrt{(1 + 2/3n)}$ (i.e., asymptotically to $1 + 1/6n$ and $1 + 1/3n$).

Having given these basic details, we can proceed rather more rapidly, arguing in terms of the more convenient modified distribution.

By distributing the mass p_h uniformly over the interval $x_h \pm 1/\sqrt{n}$, we obtain a density $f_n(x_h) = p_h / (2/\sqrt{n}) = \frac{1}{2} p_h \sqrt{n} = \frac{1}{2}\sqrt{n}\binom{n}{h} / 2^n$. By distributing it in a triangular fashion, the density at x_h remains the same, but, in every interval $[x_h, x_{h+1}]$, instead of preserving in the first and second half the values of the first and second end-points, respectively, it varies linearly (the graph = the jagged line joining the ordinates at the points $x = x_h$). In fact, the contribution of p_h decreases from x_h on, until it vanishes at x_{h+1} (and the contribution of p_{h+1} behaves in a symmetric fashion).

In the interval $x_h < x < x_{h+1}$, the derivative of the density, $f_n'(x)$, will therefore be constant:

$$f_n'(x) = (p_{h+1} - p_h) \cdot \frac{1}{2}\sqrt{n} / (2/\sqrt{n}) = \frac{1}{4} n (p_{h+1} - p_h). \tag{7.21}$$

Recalling from 7.4.2 that $p_{h+1}/p_h = (n - h)/(h + 1)$, and from the expressions for $f_n(x_h)$ and x_h that $h = \frac{1}{2}(n + x_h\sqrt{n})$ (and similarly for x_{h+1}), we have the two alternative expressions

33 In the first case, X has a uniform distribution over $|x| \leqslant 1$; in the second case, $X = X_1 + X_2$, where X_1 and X_2 are stochastically independent, and each has this uniform distribution.

$$f_n'(x) = p_h . \frac{n}{4}\left(\frac{n-h}{h+1} - 1\right) = p_h . \frac{1}{4} n \ . \frac{n-2h-1}{h+1}$$

$$= \frac{1}{2}\sqrt{n}\frac{-x_h\sqrt{n}-1}{\frac{1}{2}\left(n + x_h\sqrt{n}\right)+1} f_n(x_h) = -x_h . f_n(x_h) . \frac{1 + 1/\left(x_h\sqrt{n}\right)}{1 + \left(x_h/\sqrt{n}\right) + (2/n)},$$

and similarly,

$$f_n'(x) = p_{h+1} . \frac{1}{4} n\left(1 - \frac{h+1}{n-h}\right) = -x_{h+1} \cdot f_n(x_{h+1}) \cdot \frac{1 - 1/\left(x_{h+1}\sqrt{n}\right)}{1 - \left(x_{h+1}/\sqrt{n}\right) + (2/n)}.$$

This proves that the logarithmic derivative $f_n'(x)/f_n(x)$ (which clearly has its own extreme values to the right of the left-hand end-point, and to the left of the right-hand end-point) always satisfies the equation

$$f_n'(x)/f_n(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log f_n(x) = -x\left[1 + \varepsilon(x)\right] \tag{7.22}$$

(where, as n increases, $\varepsilon(x)$ tends uniformly to 0 in any finite interval which has a neighbourhood of the origin removed; e.g. we have $\varepsilon(x) < \varepsilon$, for some n, throughout the interval $2\varepsilon/\sqrt{n} < |x| < \sqrt{n}/2\varepsilon$; the apparent irregularity at the origin merely stems, however, from the fact that both x and $f'(x)$ go to zero, and the equation is automatically satisfied without there being any need to consider the ratio).

The limit distribution must, therefore, satisfy

$$f'(x)/f(x) = -x, \tag{7.23}$$

from which we obtain

$$\log f(x) = -\frac{1}{2}x^2 + \text{const.}, \quad f(x) = K\mathrm{e}^{-\frac{1}{2}x^2} \quad \left(K = 1/\sqrt{(2\pi)}\right). \tag{7.24}$$

The conclusion is therefore as follows: *the standardized binomial distribution* (the case of Heads and Tails, $p = \frac{1}{2}$) *tends, as* $n \to \infty$, *to the standardized normal distribution*. The same conclusion holds, however, in more general cases and, because of its importance, is known as the *central limit theorem* of the calculus of probability. We see immediately that the conclusion holds in the binomial case for $p \neq \frac{1}{2}$ (except that we now require different coefficients in order to obtain the *standardized* form).

7.6.3. It is convenient at the beginning to dwell upon the rather special example of Heads and Tails, since this provides an intuitive and straightforward illustration of many concepts and techniques, which themselves have a much broader compass, but whose essential meaning could otherwise be obscured by the technical details of the general case.

The proof we have just given (based on a technique used by Karl Pearson for this and other examples) is probably the easiest (even more so if one omits the details of the inequalities and simply makes the heuristic observation that, for large enough n, $f'(x)/f(x)$ is practically equal to $-x$).

Remark. Geometrically, this means that the *subtangent* $-1/x$ of the graph of $y = f(x)$ is inversely proportional to the abscissa. The *tail* beyond x is, approximately, an exponential distribution, with density $f(\xi) = K\,\mathrm{e}^{-x\xi}$[34] and prevision $1/x$; this is, in fact, as we can see asymptotically from equation 7.20), the prevision of the excess of X over x (provided it does exceed it). This means, essentially, that if an error X (with a standardized normal distribution) exceeds some large given value x, it is almost certain that it exceeds it by very little (about $1/x$): for example, if it exceeds 10σ (or 100σ), we can expect that it exceeds it by $\sigma/10$ (or $\sigma/100$).

Note that this is precisely what happens for the deviations of Heads and Tails (see the footnote to equation 7.4 of Section 7.4.2), provided we make appropriate allowances for the discreteness. If we know that Heads have occurred in more than 75% of the trials, the probabilities that it has occurred 1, 2, 3, 4, or more than 4 times beyond this limit are 0·67, 0·22, 0·074, 0·025, 0·012, respectively, no matter how many tosses n have been made. This means that for $n = 100$ it is almost certain that (with the probabilities given above) the number of successes is one of 76, 77, 78, 79, whereas, for $n = 1000$, the same holds for 751, 752, 753, 754, for $n = 1{,}000{,}000$, for 750,001, 750,002, 750,003, 750,004!

We shall present other (and more general) proofs of this theorem later and it will be instructive to see it tackled from different standpoints. For the moment, however, we shall consider a useful corollary of it.

Using the fact that $f_n(x) \simeq f(x)$, and recalling the relation with p_h, we find that

$$\omega_h^{(n)} = p_h \simeq \left(2/\sqrt{n}\right) f_n\left(x_h\right) \simeq \left(2/\sqrt{n}\right) f\left(x_h\right) = \sqrt{\left(2/\pi n\right)} \exp\left[-\frac{1}{2n}\left(2h-n\right)^2\right]. \qquad (7.25)$$

In particular, for $x = 0$, we obtain the maximum term, that is the central one ($h = \frac{1}{2}n$ if n = even, or either of $h = \frac{1}{2}(n \pm 1)$ if n = odd). We shall always denote this by a special symbol, u_n, and the formula we have arrived at gives the asymptotic expression $u_n \simeq \sqrt{(2/\pi n)}$; that is, in figures, $u_n \simeq 0{\cdot}8/\sqrt{n}$ (this makes clear the meaning of the coefficient $\sqrt{(2/\pi)}$, which it is important to keep in mind). In fact, the probability u_n (the maximum probability among the $\omega_h^{(n)}$ of the Heads and Tails case) will crop up in many problems (a partial summary of which will be given in Chapter 8, 8.7.4). For the present, we shall just indicate a few of its properties.

In fact, we have

$$\begin{aligned} u_n = u_{2m} \quad &= \mathbf{P}\left(Y_n = 0\right) = \omega_m^{(n)} \qquad \text{for } n = 2m = \text{even}, \\ u_n = u_{2m-1} &= \mathbf{P}\left(Y_n = 1\right) = \mathbf{P}\left(Y_n = -1\right) \\ = \omega_m^{(n)} \quad &= \omega_{m-1}^{(n)} = u_{2m} \qquad \text{for } n = 2m-1 = \text{odd}. \end{aligned} \qquad (7.26)$$

The equality of the u_n for successive pairs of values (each odd one with the next even one) is obvious from the definition. In order that the gain after $2m$ tosses be zero, it is necessary that it was either +1 or −1 at the preceding toss and that the final toss had

---

34 We have $\exp\left\{-\frac{1}{2}(x-\xi)^2\right\} = \exp(-\frac{1}{2}x^2)\exp(-x\xi)\exp(-\frac{1}{2}\xi^2)$, but only the first factor remains because the second is constant (with respect to ξ), and is incorporated into K, and the third is $\simeq 1$ (for small ξ).

the outcome required to bring it to 0; both possibilities have probability $u_{2m-1}\cdot\frac{1}{2}$, and their sum gives

$$u_{2m} = 2\left(\frac{1}{2}u_{2m-1}\right) = u_{2m-1}.$$

The same argument can be carried out for the binomial coefficients by applying Stiefel's formula. The central term, $\binom{2m}{m}$, for $n = 2m$ = even, is the sum of the two adjacent ones which are themselves equal,

$$\binom{2m-1}{m-1} = \binom{2m-1}{m},$$

and is therefore twice their value; in order to obtain the probability, however, we must divide by 2^{2m} rather than by 2^{2m-1}, so $u_{2m} = u_{2m-1}$. We obtain, therefore,

$$u_{2m-1} - u_{2m} = \binom{2m}{m} / 2^{2m} = \frac{(2m)!}{2^{2m}\left(m!\right)^2} \simeq \sqrt{(2/\pi n)} \simeq 0\cdot 8 / \sqrt{n}. \tag{7.27}$$

7.6.4. We see from this that the factor $\sqrt{(2/\pi;)}$, hitherto regarded simply as the normalization factor for the standardized normal distribution, also has a link with the combinatorial calculus. This connection is given by Stirling's formula, which provides an asymptotic expression for the factorial and which enables us to arrive at the central limit theorem for the binomial distribution by a different route (one that is more laborious but is often used and is, in any case, useful to know).

Stirling's formula expresses *n*! as follows:

$$n! = n^n \mathrm{e}^{-n} \sqrt{(2\pi n)}\left(1+\varepsilon_n\right)\left(\text{where } \varepsilon_n \to 0; \text{ more precisely, } 0 < \varepsilon_n < 1/1\ \ 1n\dagger\right)^{35}. \tag{7.28}$$

Since the formula is used so often, we shall give a quick proof of it. We have

$$\begin{aligned} \log n! &= \log 2 + \log 3 + \ldots + \log n \\ &\simeq \int_{\frac{3}{2}}^{n+\frac{1}{2}} \log x \,\mathrm{d}x = \left[x\log x - x\right]_{\frac{3}{2}}^{n+\frac{1}{2}} \\ &= \left(n+\frac{1}{2}\right)\log n - n + \text{const.}, \end{aligned} \tag{7.29}$$

and we observe that the difference between the sum and the integral converges (substituting log *n* in place of

$$\int_{n-\frac{1}{2}}^{n+\frac{1}{2}} \log x \ \mathrm{d}x,$$

35 Note that, if we neglect ε_n, Stirling's formula gives *n*! with smaller and smaller *relative* error, but with greater and greater *absolute* error (i.e. the *ratio* tends to 1, but the *difference* between *n*! and the approximation tends to $+\infty$). In practical terms, for $n \simeq 10^k$ we have *n*! with about the first $k + 1$ digits correct; but *n*! (for large *k*) has about 10^k digits, and the error has not many less. In any case, what matters in applications is the relative approximation, and this is adequate even for small values.

we see immediately that we have an error of order $1/n^2$). From this, it follows that $n! \simeq Kn^{n+\frac{1}{2}} e^{-n}$ (a result known to De Moivre). As for the fact that $K = \sqrt{(2\pi)}$ (discovered by Stirling in 1730), we shall consider it as being established heuristically by virtue of the fact that, were we to leave it indeterminate, the limit of the $f_n(x)$ would be given by

$$f(x) = (1/K)e^{-\frac{1}{2}x^2}$$

and we know that this multiplicative factor must be $1/\sqrt{(2\pi)}$.

Let us just evaluate u_n by this method (n even: $n = 2m$): we obtain

$$u_n = \binom{2m}{m} / 2^{2m} = \frac{(2m)!}{2^{2m}(m!)^2} \simeq \frac{2^{2m} m^{2m} e^{-2m} \sqrt{(2\pi 2m)}}{2^{2m}\left[m^m e^{-m} \sqrt{(2\pi m)}\right]^2}$$
$$= \frac{1}{\sqrt{(\pi m)}} = \sqrt{(2/\pi n)} = 0 \cdot 8\sqrt{n}.$$

In order to evaluate

$$\omega_{m+k}^{(n)} = \frac{(2m)!}{2^{2m}(m-k)!(m+k)!} = \omega_m^{(n)} \frac{(m!)^2}{(m-k)!(m+k)!},$$

it is more convenient to make use of an alternative form of Stirling's formula, one which will turn out to be useful in a number of other cases. It is based on evaluating products of the form $(1 + a)(1 + 2a)\ldots(1 + ka)$, with k large, and $ka = c$ small; in our case, $[m!/(m - k)!]/[(m + k)!/m!]$ can be written as

$$\left[1.(1-a)(1-2a)\ldots(1-(k-1)a)\right] / \left[(1+a)(1+2a)\ldots(1+ka)\right]$$

by dividing both ratios by m^k, and setting $1/m = a$.

Taking the logarithm, we have

$$\log \prod_{h=1}^{k} (1+ah) = \sum_{h=1}^{k} \log(1+ah) \simeq \frac{1}{a} \int_1^{\lambda} \log x \, dx$$
$$= \frac{1}{a}\left[(1+\lambda)\log(1+\lambda) - \lambda\right],$$

with $\lambda = (k + \frac{1}{2})a$,[36] and, expanding in a series, we have

$$\log \prod_{h=1}^{k} (1+ah) = \frac{\lambda^2}{2a}\left(1 - \frac{\lambda}{3} + \frac{\lambda^2}{6} - \frac{\lambda^3}{10} + \ldots \pm \frac{\lambda_n}{\frac{1}{2}(n+1)(n+2)} \mp \ldots\right)$$
$$\simeq \frac{1}{2}\left(k + \frac{1}{2}\right)^2 a.$$

36 The simpler form $\lambda = ka$ is practically equivalent to the effect of an individual evaluation. In the case of products or ratios of a number of expressions of this kind, however, it can happen (and does in the example of Section 7.4.3) that it is the contributions deriving from the '+½' which are important, because the main contributions cancel out.

It follows that

$$(1+a)(1+2a)\ldots(1+ka) \simeq e^{\frac{1}{2}\left(k+\frac{1}{2}\right)^2 a} \simeq e^{\frac{1}{2}ak^2}. \tag{7.30}$$

In our case, with $a = \pm 1/m$, the two products equal $e^{\pm\frac{1}{2}ak^2} = e^{\pm k^2/2m}$ and their ratio is

$$e^{-k^2/2m} / e^{k^2/2m} = e^{-k^2/m} = e^{-(h-m)^2/m} = e^{-(2h-n)^2/2n}$$

(since $k = h - m$ and $m = \frac{1}{2}n$). We thus obtain the result (which, of course, we already knew).

7.6.5. *Relation to the diffusion problem.* We give here a suggestive argument (due to Pólya), which is entirely heuristic, but is very useful as a basis for discussions and developments. The relation between random processes of the kind we have just exemplified with Heads and Tails and diffusion processes, which we shall meet later, will, in fact, provide a basis for interpreting the latter and even identifying them with the kinds of process already studied. The Wiener–Lévy process (see Chapter 8) can, with reference to our previous work, be thought of as a Heads and Tails process involving an enormous number of tosses with very small stakes, taking place at very small time intervals. This process has also been referred to (by P. Lévy) as the *Brownian motion* process, because it can be used (although only for certain aspects of the problem) to represent and study the phenomenon of the same name (which is, as is well known, a diffusion process).

The Heads and Tails process can be thought of as a diffusion process in which a mass (a unit mass, initially – i.e. at $t = 0$ – concentrated at the origin) moves, with respect to time t, through the lattice of Figure 7.2, splitting in half at each intersection (encountered at times t = integer). The mass (which represents the probability) would, according to this representation, divide up in a certain (i.e. deterministic) manner, and, formally, everything goes through (indeed, it will be even simpler than this).

A more meaningful interpretation, however, and one more suited to our purpose, derives from consideration of a random process of the statistical type. Assume that, initially, a very large number of particles (N, say) are concentrated at the origin, and move at equal and constant rates to the right, through the lattice. At each time instant t = integer, they meet an intersection, and each chooses its direction independently of the others. Equivalently, we could think of them as moving with constant speed on the y-axis, choosing directions at random at each time instant t = integer (i.e. each time a point y = integer is reached); alternatively, we could think of them at rest, but making a jump of ±1 at each t = integer.

Taking the total mass = 1, the mass crossing a given point can no longer be determined with certainty: where, in the deterministic case, it was ω, we can now only say that we have prevision ω and that the number of particles has prevision $N\omega$, but could take any value h, lying between 0 and N, with probability $\binom{N}{h}\omega^h(1-\omega)^{N-h}$. If we want to give a rough idea of what happens, we could say (quoting the prevision ± the standard deviation) that the number of particles will be

$$N\omega \pm \sqrt{\left[N\omega(1-\omega)\right]}$$

($\simeq N\omega \pm \sqrt{(N\omega)}$ for small ω; the Poisson approximation).

This is what we are interested in: a normal distribution being attained as a result of a statistical diffusion process.

For the purposes of the mathematical treatment (whatever the interpretation), the mass crossing the point (vertex) (t, y), where t and y are integer, both even or both odd, is, as usual,

$$\mathbf{P}\left(Y_t = y\right) = \omega_{(t+y)/2}^{(t)} = \omega\left(t, y\right),$$

given by one half of that which has crossed $(t - 1, y - 1)$ or $(t - 1, y + 1)$:

$$\omega\left(t,\ y\right) = \frac{1}{2}\left[\omega\left(t-1,\ y-1\right) + \omega\left(t-1,\ y+1\right)\right].$$

The notation $\omega(t, y)$ has been introduced in order to allow us to think of the function as defined everywhere (no matter what the interpretation), even on those points where it has no meaning in the actual problem; in particular, for t and y integer, but $t + y$ odd, like $\omega(t - 1, y)$. Subtracting this value from both sides of the previous equation, we obtain

$$\Delta_t \omega = \frac{1}{2}\Delta_y^2 \omega \tag{7.31}$$

and, in the limit,

$$\frac{\partial \omega}{\partial t} = \frac{1}{2}\frac{\partial^2 \omega}{\partial y^2}, \tag{7.32}$$

provided that (taking the units of t and y to be very small) one considers it legitimate to pass from the discrete to the continuous.

Let us restrict ourselves here to simply pointing out that, in this way, one arrives at the correct solution. In fact, the general solution of the *heat equation*, (7.32), is given by

$$\omega\left(t, y\right) = \left(K/\sqrt{t}\right)\mathrm{e}^{-\frac{1}{2}y^2/t}, \tag{7.33}$$

a well-known result that can easily be verified.

7.6.6. The form of the normal distribution is well known, and is given in Figure 7.6 (where we show the density $y = f(x)$). We also provide a table of numerical values for both the density and the distribution function (the latter giving the probabilities of belonging to particular half-lines or intervals).

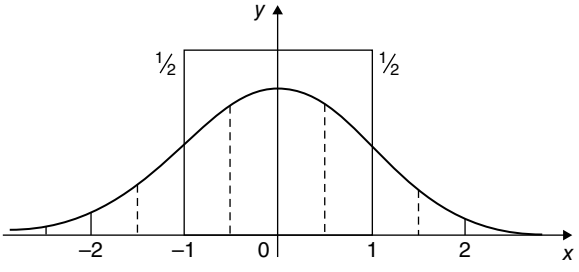


Figure 7.6 The standardized normal distribution ($m = 0$, $\sigma = 1$): the density function. The subdivisions $(0, \pm1, \pm2, \pm3)$ correspond to σ, 2σ, 3σ; at ±1 we have points of inflection, between which the density is convex. The rectangle of height $\frac{1}{2}$ shows, for comparative purposes, the uniform distribution on the interval $[-1, +1]$ (which might well be called the 'body' of the distribution; see Chapter 10, 10.2.4). Note that the vertical scale is, in fact, four times the horizontal one, in order to avoid the graph appearing very flat (as it is, in fact), and hence not displaying the behaviour very clearly.

We shall confine ourselves to calling attention to a few points of particular importance.

The density function is symmetric about its maximum, which is at the origin, and decreases away from it, being convex (upwards) in the interval [−1, +1], and concave outside this interval. As $x \pm \infty$, it approaches the x-axis, the approach being very rapid, as we already pointed out in 7.6.3 (the 'tails' are 'very thin'). In fact, the subtangent decreases, in our case, like $1/|x|$ (if $f(x)$ tends to 0 like a power, $|x|^{-n}$, with n arbitrarily large, the subtangent, in absolute value, increases indefinitely as $|x|$ increases; if the function decreases exponentially, the subtangent is constant).

The graph has two points of inflection at ±1 (corresponding to the change from convexity to concavity that we already mentioned). The ordinate at these points is about 0·6 of the maximum value (the table gives 0·60652, but it is better to keep an approximate round figure in mind; this is enough to prevent one from making the all too usual distortions when sketching it – on the blackboard, for example). The subtangents are equal to ∓ 1; that is the slope is such that the tangents cross the x-axis at the points ±2.

Since the tails are 'very thin', it is clear that the probabilities of the occurrence of extreme values beyond some given x are, in the case of the normal distribution, much smaller than is usual (in the case of densities decreasing like powers, or exponentially). They are, therefore, much smaller than the values provided by Tchebychev's inequality, which is valid under very general conditions.

We give below a few examples of the probabilities of $|X|$ exceeding $k\sigma$ (or, in the standardized case, $\sigma = 1$, of exceeding k), for $k = 1, 2, 3$ and $3\frac{1}{2}$:

Absolute value greater than:		σ	2σ	3σ	$3\frac{1}{2}\sigma$
Probability	normal distribution:	31·74%	4·55%	0·27%	0·05%
	Tchebychev inequality:	⩽100·00%	⩽25·00%	⩽11·11%	⩽8·16%

The table is not only useful for numerical applications but it should also be used in order to commit to memory a few of the significant points (e.g. a few ordinates and, more importantly, the areas corresponding to abscissae 1, 2 and 3; i.e. to σ, 2σ and 3σ).

The reader is invited to refer to equation 7.20 in Section 7.5.4, and to the *Remarks* of Section 7.6.3, where we looked at asymptotic expressions for such probabilities ($\simeq K\mathrm{e}^{-\frac{1}{2}x^2}/x$, $K = 1/\sqrt{(2\pi)} \simeq 0{\cdot}40$[37]), and at the order of magnitude for possible exceedances (prevision $\simeq 1/x$).

Table of values for the standardized normal (Gaussian) distribution

Abscissa	Ordinate (density)		Area ($\int f(x)\,dx$) in %		
X	$f(x) = \frac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{1}{2}x^2}$	$f(x)$ as % of central ordinate	from x to $+\infty$	in the individual intervals given	2 × (5)
(1)	(2)	(3)	(4)	(5)	(6)
0·0	0·398942	100·0	50·0		
0·1	0·396952	99·50	46·0172		
0·2	0·391043	98·02	42·0740	19·15	38·30
0·3	0·381388	95·60	38·2089		
0·4	0·368270	92·31	34·4978		
0·5	0·352065	88·25	30·8538		

37 Writing $K(1 + \Theta/x^2)$, with $0 \leqslant \Theta \leqslant 1$, in place of K, we have an exact bound (and $\Theta \sim 1 - 3/x^2$).

Abscissa	Ordinate (density)		Area ($\int f(x)\,dx$) in %		
X	$f(x)=\frac{1}{\sqrt{(2\pi)}}e^{-\frac{1}{2}x^2}$	$f(x)$ as % of central ordinate	from x to $+\infty$	in the individual intervals given	2 × (5)
(1)	(2)	(3)	(4)	(5)	(6)
0·6	0·333225	83·53	27·4253		
0·7	0·312254	78·27	24·1964		
0·8	0·289692	72·61	21·1855	14·98	29·96
0·9	0·266085	66·70	18·4060		
1·0	0·241971	60·652	15·8654		
1·1	0·217852	54·61	13·5666		
1·2	0·194186	48·68	11·5070		
1·3	0·171369	42·96	9·6800	9·19	18·38
1·4	0·149727	37·53	8·0756		
1·5	0·129518	32·47	6·6807		
1·6	0·110921	27·80	5·4799		
1·7	0·094049	23·57	4·4565		
1·8	0·078950	19·79	3·5930	4·405	8·810
1·9	0·065616	16·45	2·8717		
2·0	0·053991	13·53	2·2750		
2·5	0·017528	4·39	0·6210	2·140	4·280
3·0	0·004432	1·11	0·1350		
3·5	0·0008727	0·22	0·02326	0·135	0·270
∞	0·0	0	0·0		

Since the binomial distribution (and many others) approximates, under given conditions, as n increases to the normal, the table of the latter can also be used in other contexts (with due care and attention[38]). Such tables are often used in the case of empirical distributions (statistical distributions), under the confident assumption that the latter behave (at least approximately) like normal distributions. It is easily shown (see Section 7.6.9) that this confidence is often not, in fact, justified.

7.6.7. In order to deal with certain other instructive and important features of the normal distribution, we shall have to refer to the multidimensional case (either just two dimensions, the plane, or some arbitrary number n; or even the asymptotic case, $n \to \infty$).

It will suffice to limit our discussion to the case of spherical symmetry, where the density has the form

$$f\left(x_1, x_2, \ldots, x_r\right) = K \exp\left(-\frac{1}{2}\rho^2\right), \quad \rho^2 = x_1^2 + x_2^2 + \ldots + x_r^2.$$

This corresponds to assuming the X_h to be standardized ($m = 0, \sigma = 1$) and stochastically independent (for which, in the case of normality, it is sufficient that they be uncorrelated). In fact, we can always reduce the general situation to this special case provided we apply to S_r the affine transformation that turns the covariance ellipsoid into a 'sphere' (see Chapter 4, 4.17.6). In other words, we change from the X_h to a set of Y_k which are

38 If one were not careful, one might conclude that the probability of obtaining more than n Heads in n tosses(!) is very small, but not zero (about $2{\cdot}4 \times 10^{-23}$ for $n = 100$; about 10^{-2173} for $n = 10,000$).

standardized and uncorrelated (and are linear combinations of the X_h). We shall have more to say about this later (Chapter 10, 10.2.4).

For the moment, let us evaluate the normalizing constant of the standardized normal distribution (which we have already stated to be $K = 1/\sqrt{(2\pi)}$). Integrating over the plane, we obtain

$$\iint \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}x\,\mathrm{d}y = \int \mathrm{e}^{-\frac{1}{2}\rho^2}\,\rho\,\mathrm{d}\rho \int \mathrm{d}\theta = 2\pi,$$

which is also equal to $\left[\int \mathrm{e}^{-\frac{1}{2}x^2}\mathrm{d}x\right]^2$. It follows that $K = 1/2\pi$ in the plane ($r = 2$), $K = (2\pi)^{-\frac{1}{2}}$ over the real line ($r = 1$), and, for general r, $K = (2\pi)^{-\frac{1}{2}r}$.[39]

There are another two important, interesting properties to note. They involve the examination of two conditions, closely linked with one another, each of which provides a meaningful characterization of the normal distribution. In both cases, it is sufficient to deal with the case of the plane.

The first of them is summarized in the following: the only distribution over the plane which has circular symmetry, and for which the abscissa X and the ordinate Y are stochastically independent (orthogonal), is that in which X and Y have normal distributions with equal variances (and zero prevision, assuming the symmetry to be about the origin). The second property concerns the *stability* of distributions (which we discussed in Chapter 6, 6.11.3). If we require, in addition, that the variance be finite, then stability is the *exclusive* property of the normal distribution.

For the first property, if we denote by $f(\cdot,\cdot)$, $f_1(\cdot)$, $f_2(\cdot)$ the joint density for (X, Y), and the marginal densities for X and Y, respectively, the given conditions may be expressed as follows:

a) *rotational symmetry;* $f(x, y) = \text{const}$, for $x^2 + y^2 = \rho^2 = \text{const.}$, from which it follows immediately that $f(x, y) = f(\rho, 0) = f(0, \rho)$ for $\rho = \sqrt{(x^2 + y^2)}$;
b) independence; $f(x, y) = f_1(x)f_2(y)$.

In our case, given the symmetry, we can simply write $f(\cdot)$ instead of $f_1(\cdot)$ and $f_2(\cdot)$, and hence obtain a single condition,

$$f(x, y) = f(x)f(y) = f(0)f(\rho)$$

In other words,

$$\frac{f(x)}{f(0)}\frac{f(y)}{f(0)} = \frac{f(\rho)}{f(0)},$$

and, if we put $f(x)/f(0) = \psi(x^2)$, this gives the functional equation

$$\psi(x^2)\psi(y^2) = \psi(\rho^2) = \psi(x^2 + y^2).$$

39 We should make it clear (because it is customary to do so – it is, in fact, obvious) that the integral of a positive function taken over the plane does not depend on how one arrives at the limit (by means of circles, $\rho < R$, or squares, $|x| \vee |y| < R$, or whatever); it is always the supremum of the values given on the bounded sets.

Taking logarithms, this gives the additive form

$$\log\psi\left(x^2+y^2\right)=\log\psi\left(x^2\right)+\log\psi\left(y^2\right).$$

Under very weak conditions, which are usually satisfied, this implies linearity (e.g. it is sufficient that ψ is non-negative in the neighbourhood of some point; here, this holds over the whole positive real line). It follows that

$$\log\psi\left(x^2\right)=kx^2,\quad \psi\left(x^2\right)=\mathrm{e}^{kx^2},\quad f(x)=f(0)\mathrm{e}^{kx^2}$$

(and, normalizing, that $k=-1/2\sigma^2$ and $f(0)=1/\sqrt{(2\pi)}\sigma$); the required property is therefore established.

In certain cases, this property is in itself sufficient to make the assumption of a normal distribution plausible. A celebrated example is that of the distribution of the velocity of the particles in Maxwell's kinetic theory of gases. If one assumes: (a) *isotropy* (the same distribution for components in all directions) and (b) stochastic *independence* of the orthogonal components, then the distribution of each component is normal (with zero previsions and equal variances). In other words, the distribution of the velocity vector is normal and has spherical symmetry (with density $K\mathrm{e}^{-\frac{1}{2}\rho^2/\sigma^2}$).

Given the assumptions, the above constitutes a mathematical proof. But however necessary they are as a starting point, the question of whether or not these (or other) assumptions should be taken for granted, or regarded as more or less plausible, is one which depends in part upon the actual physics, and in part upon the psychology of the author concerned.

The second property referred to above reduces to the first one. We must first of all restrict ourselves to the finite variance case (otherwise, we already know the statement to be false; see the stable, Cauchy distribution, mentioned at the end of Chapter 6, 6.11.3), and we might as well assume unit variance. We therefore let $f(x)$ denote the density of such a distribution (with $m = 0$, $\sigma = 1$), and X and Y be two stochastically independent random quantities having this distribution.

In order for there to be stability, $Z = aX + bY$ must, by definition, have the same distribution (up to a change of scale, since $\sigma^2 = a^2 + b^2$). If, by taking $a^2 + b^2 = 1$, we make $Z = X\cos\alpha + Y\sin\alpha$, we can avoid even the change of scale, and we can conclude that all projections of the planar distribution, $f(x, y) = f(x)f(y)$, in whatever direction, must be the same. In other words, the projections must possess circular symmetry and it can be shown that a necessary condition for this (and clearly a sufficient one also) is that the density has circular symmetry (as considered for the first property).[40]

The conclusion is, therefore, the same: the property characterizes the normal distribution. The result contains within it an implicit justification (or, to be more accurate, a partial justification) of the 'central limit theorem'. In fact, if the distribution (standardized, with $\sigma = 1$) of the gain from a large number of trials at Heads and Tails follows, in practice, some given distribution, then the latter must be stable (and the same is true for any other case of stochastically independent gains). It is sufficient to note that if Y' and Y'' are the gains from large numbers of trials, n' and n'', respectively, then, *a fortiori*,

40 This is intuitively entirely 'obvious'. The proof, which is rather messy if one proceeds directly, follows immediately from the properties of characteristic functions of two variables (Chapter 10, 10.1.2).

$Y = Y' + Y''$ is the gain from $n = n' + n''$ trials. If the two independent summands belong to the limit distribution family, then so does their sum: this implies stability.

The justification is only partial because the above argument does not enable us to say whether, and in which cases (not even for that of Heads and Tails), there is convergence to a limit distribution. It does enable us to say, however, that *if a limit distribution exists* (with a finite standard deviation, and if the process is additive with independent summands – all obvious conditions) *it is necessarily the normal distribution*.

7.6.8. *An interpretation in terms of hyperspaces*. It is instructive to bear in mind, as an heuristic, but meaningful, interpretation, that which can be given in terms of hyperspaces. Compared with the previous example, it constitutes even less of a 'partial justification' of the appearance of the normal distribution under the conditions of the central limit theorem (sums of independent random quantities), but, on the other hand, it reveals how the result is often the same, even under very different conditions.

Let us begin by considering the uniform distribution inside the sphere (hypersphere) of unit radius in S_r, and the projection of this distribution onto the diameter, $-1 \leqslant x \leqslant +1$. The section at x has radius $\sqrt{(1-x^2)}$, 'area' equal to $[\sqrt{(1-x^2)}]^{r-1}$, and hence the density is given by

$$f(x) = K\left(1-x^2\right)^{(r-1)/2}. \tag{7.34}$$

In particular, we have $K\sqrt{(1-x^2)}$ for $r = 2$ (projection of the area of the circle); $K(1-x^2)$ for $r = 3$ (projection of the volume of the sphere); and so on.

As r increases, the distribution concentrates around the origin (as happened in the case of frequencies at Heads and Tails). In order to avoid this and to see what, asymptotically, happens to the 'shape' of the distribution, it is necessary (again, as in the case of Heads and Tails) to expand it in the ratio 1: $\sqrt{r}$ (i.e. by replacing x by $x/\sqrt{r}$). We then obtain

$$f(x) = K\left(1-\frac{x^2}{r}\right)^{(r-1)/2} \to K\mathrm{e}^{-\frac{1}{2}x^2}.$$

In the limit, this gives the normal distribution, but without any of the assumptions of the central limit theorem. What is more surprising, however, is that the same conclusion holds under circumstances even less similar to the usual ones. For example, it also holds if one considers a hollow sphere, consisting of just a small layer between $1-\varepsilon$ and 1 ($\varepsilon > 0$ arbitrary). It is sufficient to note that the mass inside the smaller sphere contains $(1-\varepsilon)^r$ of the total mass. This tends to 0 as r increases, and hence its contribution to the determination of the shape of $f(x)$ becomes negligible.

Well: the central limit theorem, also, can be seen as a special case of, *so to speak*, this kind of tendency for distributions in higher dimensions to have normally distributed projections onto a certain straight line.

The case of Heads and Tails shows that one obtains this projection (asymptotically, for large n) by projecting (onto the diagonal) a distribution of equal masses $\left(\frac{1}{2}\right)^n$ on the 2^n vertices of an n-dimensional hypercube. The same holds, however, for projections onto any other axis (provided it does not belong entirely to a face having only a small number of dimensions compared with n) and also if one thinks of the cube as a solid (with uniformly distributed mass inside it), or with a uniform distribution on the

surface, and so on. To summarize; the interpretation in terms of hyperspaces holds in all cases where the central limit theorem holds (although it cannot be of any help in picking out these cases, except in those cases where there exists an heuristic argument by analogy with some known case).

More specifically useful is the conclusion that can be drawn in the opposite sense: namely, that of convergence to the normal distribution in many more cases than those, already numerous, which fall within the ambit of sums of independent random quantities, the case we are now considering (having started with the case of Heads and Tails). The solid cube does fall within such an interpretation (summands chosen independently and with a uniform distribution between ±1), but that of a distribution on the surface (or on edges, or m-dimensional faces) does not, and this would be even less true for the case of the hypersphere (solid, or hollow).

A wide-ranging generalization of the central limit theorem was given by R. von Mises,[41] and shows that even the distributions of nonlinear 'statistical functions' may be normal (under conditions which, in practice, are not very restrictive). Examples of nonlinear statistical functions of the observed values $X_1, X_2,\ldots, X_n$ are the means (other than the arithmetic mean, or their deviations from it), the moments and the functions of the moments (e.g. μ_3^2/μ_2^3, or $(\mu_4/\mu_2^2)-3$, where the μ_h are the moments about the mean and the expressions are used as indices of asymmetry and 'kurtosis', respectively; see Chapter 6, 6.6.6), the concentration coefficient (of Gini; see Chapter 6, the end of 6.6.3), and so on. In general, they are the functions that can be interpreted as functionals of the $F_n(x) = (1/n)\,\Sigma(X_h \leqslant x)$ (jump $1/n$ for $x = X_1, X_2,\ldots, X_n$), that is of the statistical distribution, under conditions similar to differentiability (i.e. of local linearity'). In essence (the actual formulation is quite complicated and involves long preliminary explanations before one can even set up the notation), one requires that the first derivative (in the sense of Volterra, for 'fonctions de ligne') satisfies the conditions for the validity of the 'central limit theorem' in the linear case and that the second derivative satisfies a complementary condition.

This generalization, wide ranging though it is, does not, however, include the cases that we considered in the hyperspace context. This emphasizes even further just how general is the 'tendency' for the normal distribution to pop up in any situation involving 'chaos'.

7.6.9. *Order out of chaos*. We shall postpone what we consider to be a *valid* proof of the central limit theorem until the next section (from a mathematical viewpoint it is a stronger result). Let us consider first the notion of 'order generated out of chaos', which has often been put forward in connection with the normal distribution (as well as in many other cases).

A general observation, which is appropriate at this juncture, concerns a phenomenon that often occurs in the calculus of probability; that of obtaining conclusions which are extremely precise and stable, in the sense that they hold unchanged even when we start from very different opinions or situations. This is the very opposite of what happens in other fields of mathematics and its applications, where errors pile up and have a cumulative effect, with the risk of the results becoming completely invalidated, no matter

41 R. von Mises, *Selected Papers*, Vol. II, Providence (1964); see various papers, among which (pp. 388–394) the lectures given in Rome (Institute of Advanced Mathematics) provide one of the most up to date expositions (for an exposition of a more illustrative kind, see pp. 246–270).

how carefully the initial data were evaluated and the calculations carried out. This particular phenomenon compensates for the disadvantages inherent in the calculus of probability due to the subjective and often vague nature of the initial data. It is because of this peculiarity (which is, in a certain sense, something of a miracle, but which, after due consideration, can be seen, in a certain sense, as natural) that a number of conclusions appear acceptable to everyone, irrespective of inevitable differences in initial evaluations and opinions. This is a positive virtue, notwithstanding the drawbacks which stem from a too indiscriminate interpretation of it, leading one to accept as objective those things whose roots are, in fact, subjective, but have not been explicitly recognized as such.

In this connection, we shall now put forward an example that is basically trivial but, nonetheless, instructive (because it is clear what is going on; the central limit theorem is less self-evident). We return to case (*E*) of Section 7.2.2 and consider the probability of an odd number of successes out of n events, $E_1, E_2, \ldots, E_n$. If we assume them to be stochastically independent with probabilities $p_1, p_2, \ldots, p_n$, the probability in question is given by

$$q_n = \frac{1}{2} - \prod_{h=1}^{n} (1 - 2p_h)$$

(which can be verified by induction). As n increases, the difference between q_n and $\frac{1}{2}$ decreases (in absolute value). In other words, if one is interested in obtaining a probability close to $\frac{1}{2}$, it is always better to add in (stochastically independent) events, whatever the probabilities p_h might be, because the above-mentioned difference is multiplied by $2(\frac{1}{2} - p_h)$, which is $\leqslant 1$ in absolute value, and the smaller the difference is, the closer p_h is to $\frac{1}{2}$: if $p_h = \frac{1}{2}$, the difference becomes zero (as we remarked at the time). Suppose we now consider a cube or a parallelepiped that we wish to divide into two equal parts as accurately as possible. Making use of the above, instead of performing only one cut (parallel to a face) we could perform three cuts (parallel to the three pairs of faces) and then make up a half with the four pieces that satisfy one or all of the three conditions of being 'above', 'in front', or 'on the left' (and the other half with the four pieces satisfying two or none of these conditions). (What is the point of these digressions? They are an attempt to show that phenomena of this kind do not derive from the principles or assumptions of probability theory – in which case one might well have called them 'miraculous'. They may show up, in their own right, in any kind of applications whatsoever. The fact is simply that *the exploitation and the study* of methods based on disorder is more frequent and 'relevant' in probability theory than elsewhere.)

This should give some idea (as well as, in a sense, some of the reasons) of why it is, in complicated situations where some kind of 'disorder' prevails, that something having the appearance of 'order' often emerges.

A further fact, which serves to 'explain' why it is that this 'order generated out of chaos' often has the appearance of a normal distribution, is that out of all distributions having the same variance the normal has maximum entropy (i.e. the minimum amount of information).

Among the discrete distributions with preassigned possible values x_h and prevision $\sum p_h x_h = m$, those which maximize $\sum p_h |\log p_h|$ (the sums $\sum p_h = 1, \sum p_h x_h$ and $\sum p_h x_h^2$ being fixed) are obtained by setting the $\partial/\partial p_h$ of $\sum p_h \{|\log p_h| + Q(x)\}$ equal to 0, where

$Q(x)$ is a second degree polynomial. In other words, we set $-\log p_h + Q(x) = 0$, from which it follows that

$$p_h = \exp\bigl(-Q(x)\bigr) = K \exp\left\{-\frac{1}{2}(x-m)^2\right\}.$$

If we choose the x_h to be equidistant from each other, and then let this distance tend to zero, we obtain the normal distribution.

In Chapter 3, 3.8.5, we briefly mentioned the idea of information but without going at all deeply into it. In the same way here, without going into the relevant scientific theory, we merely note that, when considering the distribution of velocities in the kinetic theory of gases, 'the same variance' corresponds to 'kinetic energy being constant' (and this may suggest new connections with Maxwell's conclusions; see Section 7.6.7, above).

We could continue in a like manner: there appear to be an endless variety of ways in which the tendency for the normal distribution to emerge occurs.[42]

It is easy to understand the wonder with which its appearance in so many examples of statistical distributions (e.g. in various characteristics of animal species etc.) was regarded by those who first came across the fact, and it is also easy to understand the great, and somewhat exaggerated, confidence in its universal validity that followed.

A typical expression of this mood is found in the following passage of Francis Galton's (it appears in his book *Natural Inheritance*, published in 1889, in the chapter entitled, 'Order in apparent chaos', and the passage is reproduced by E.S. Pearson in one of his 'Studies in the history of probability and statistics' (*Biometrika* (1965), pp. 3–18), which also contains a number of other interesting and stimulating quotations):

> 'I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. The tops of the marshalled row form a flowing curve of invariable proportions; and each element, as it is sorted into place, finds, as it were, a pre-ordained niche, accurately adapted to fit it. If the measurement of any two specified Grades in the row are known, those that will be found at every other Grade, except towards the extreme ends, can be predicted in the way already explained, and with much precision.'

Are statements of this kind acceptable? It seems to me the answer can be both yes and no. It depends more on the nuances of interpretation than on any general principle of whether such statements are correct or not.

---

42 A similar 'tendency' for the normal distribution to appear operates, although in a different manner, in problems of statistical inference, as a result of more and more information being acquired. We mention this now merely to make the above survey complete, and in no way to anticipate what will be said later (Chapter 11, 11.4.6–11.4.7, and Chapter 12, 12.6.5).

The idea that all natural characteristics have to be normally distributed is one that can no longer be sustained: it is a question that must be settled empirically.[43] What we are concerned with in the present context, however (and not only in relation to the passage above but also to the numerous other statements, of a more or less similar nature, that one comes across practically everywhere), are the attitudes adopted in response to the 'paradox' of a 'law' governing the 'accidental', which surely obeys no rules.

Perhaps the following couple of sentences will suffice as a summary of the circumstances capable of differentiating and revealing the attitudes which I, personally, would characterize as 'distorted' or 'correct', respectively:

a) there exist chance phenomena which are really under control, in that they follow the 'rules of chance phenomena', and there are others which are even more chancy, accidental in a more extreme sense, irregular and unforeseeable, occurring 'at random', without even obeying the 'laws of chance phenomena';
b) chance phenomena – the completely accidental, those which are to a large extent irregular or unforeseeable, those occurring 'at random' – are those which presumably 'obey the laws of chance phenomena'; these laws express no more and no less than that which can be expected in the absence of any factor which allows one to a large extent to foresee something falling outside the ambit formed by the overwhelming majority of the vast number of possible situations of chaos.

Even when expressed in this way, the two alternatives are very vague (and it would be difficult to avoid this – I certainly did not succeed). They may be sufficient, however, to remove some of the ambiguity from Galton's position, because they show up what the essential ambiguity is that has to be overcome.

Having said this, it remains for me to make clear that I consider (*a*), the *first* interpretation, to be '*distorted*', and (*b*), the *second* interpretation, to be the *correct one*.

The reasons for this are those that have been presented over and over again in the context of concrete problems. There is no need to repeat them here, nor is there any point in adding further general comments or explanations; these, I am afraid, would inevitably remain at a rather vague level.

7.7 Proof of the Central Limit Theorem

7.7.1. We now give the proof of the central limit theorem. This is very short if we make use of the method of *characteristic functions* – although this has the disadvantage of operating with purely analytic entities, having nothing to do with one's intuitive view of the problem. It has the advantage, however, that the very simple proof that can be given for the case of Heads and Tails (confirming something that we have already established in a variety of alternative ways) will turn out to be easily adapted, with very little effort, to provide a proof for very much more general cases.

43 One must not adopt the exaggerated view that all, or almost all, statistical distributions are normal (a habit which is still widespread, although not so much as it was in the past). Around 1900, Poincaré made the acute observation that 'everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact'.

For the gain at a single trial of Heads and Tails ($X_i = \pm 1$ with $p_i = \frac{1}{2}$ the characteristic function is given by

$$\frac{1}{2}\left(e^{iu} + e^{-iu}\right) = \cos u.$$

For the sum Y_n of n such trials (stochastically independent summands), we have $(\cos u)^n$. In the standardized form, $Y_n/\sqrt{n}$, this becomes $[\cos(u/\sqrt{n})]^n$, with logarithm equal to $n \log \cos(u/\sqrt{n})$.

Since $\log \cos x = -\frac{1}{2}x^2[1+\varepsilon(x)]$ (where $\varepsilon(x) \to 0$ as $x \to 0$), we have $n \log \cos(u/\sqrt{n}) = -\frac{1}{2}n(u/\sqrt{n})^2[1+\varepsilon(u/\sqrt{n})] = -\frac{1}{2}u^2[1+\varepsilon(u/\sqrt{n})] \to -\frac{1}{2}n^2$, and, passing from the logarithm back to the characteristic function, we obtain

$$\left[\cos\left(u/\sqrt{n}\right)\right]^n \to e^{-\frac{1}{2}u^2} \quad \text{as } n \to \infty. \tag{7.35}$$

This is precisely the characteristic function of the standard normal distribution, and so the theorem is proved.

But this does not merely hold for the case of Heads and Tails. The essential property that has been used in the proof is not the fact that the characteristic function of the individual gain is given by $\varphi(u) = \cos u$, but only that its qualitative behaviour in the neighbourhood of the origin is

$$\log \phi(u) = -\frac{1}{2}u^2\left[1+\varepsilon(u)\right].$$

This requires only that the variance be finite (the value 1 is merely due to the convention adopted previously).

Therefore: *the central limit theorem holds for sums of independent, identically distributed random quantities provided the variance is finite.*[44]

It is clear, however, that the conclusion does not require the distributions to be identical, nor the variances to be equal: given the qualitative nature of the circumstances which ensure the required asymptotic behaviour, purely qualitative conditions should suffice.

It is perhaps best to take one step at a time, in order to concentrate attention on the two different aspects separately. Let us begin with the assumption that the distributions do not vary but that the variances may differ from trial to trial (to be accurate, we should say that the type of distribution does not vary; for the sake of simplicity, we shall continue to assume the prevision to be zero).

---

44 If the variance is infinite, the central limit theorem can only hold in what one might call an anomalous sense; that is by not dividing Y_n by $\sqrt{n}$, as would be the case for the normal distribution itself, but rather, if at all, through some other kind of standardization procedure, $(Y_n \square A_n)/B_n$, with A and B appropriate functions of n. This holds (see Lévy, *Addition*, p. 113) for those distributions for which the mass outside $\pm x$, if assumed concentrated at these points, has a moment of inertia about the origin which is negligible compared to that of the masses within $\pm x$ (i.e. the ratio tends to zero as $x \to \infty$). These distributions, plus those with finite variances, constitute the 'domain of attraction' of the normal distribution.

There exist other stable distributions (with infinite variances), each having its own domain of attraction (see Chapter 8, Section 8.4).

In other words, we continue to consider the X_i to be independently and identically distributed, standardized random quantities ($\mathbf{P}(X_i) = 0$, $\mathbf{P}(X^2) = 1$), but with the summands X_i replaced by $\sigma_i X_i$ (with $\sigma_i > 0$, and varying with i). Explicitly, we consider the sums

$$Y_n = \sigma_1 X_1 + \sigma_2 X_2 + \ldots + \sigma_n X_n.$$

We again let $\phi(u)$ denote the characteristic function of the X_i and $\varepsilon(u)$ the correction term defined by $\log \phi(u) = -\frac{1}{2}u^2(1 + \varepsilon(u))$. The characteristic function of $\sigma_i X_i$ is then given by $\phi(\sigma_i u)$, with

$$\log \phi(\sigma_i u) = -\frac{1}{2}u^2\left[\sigma_i^2 + \sigma_i^2 \varepsilon(\sigma_i u)\right].$$

By taking the product of the ϕ, and the sum of the logarithms, we obtain, for the sum Y_n,

$$\begin{aligned} \log \prod_{i=1}^{n} \phi(\sigma_i u) &= \sum_{i=1}^{n} \log \phi(\sigma_i u) = -\frac{1}{2}u^2\left[\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}\sigma_i^2 \varepsilon(\sigma_i u)\right] \\ &= -\frac{1}{2}s_n^2 u^2\left[1 + \sum_{i=1}^{n}\frac{\sigma_i^2 \varepsilon(\sigma_i u)}{s_n^2}\right], \end{aligned} \tag{7.36}$$

where s_n^2 denotes the variance of Y_n; that is

$$s_n^2 = \mathbf{P}(Y_n^2) = \sum_{i=1}^{n}\sigma_i^2.$$

For the standardized Y_n, that is Y_n/s_n, we have (substituting u/s_n for u)

$$-\frac{1}{2}u^2\left[1 + \sum_{i=1}^{n}\frac{\sigma_i^2}{s_n^2}\varepsilon\left(\frac{\sigma_i u}{s_n}\right)\right], \tag{7.37}$$

and, hence, the validity of the central limit theorem depends on the fact that the 'correction term', given by the sum, tends to 0 as $n \to \infty$.

The sum in question is a weighted mean (with weights σ_i^2) of the $\varepsilon(\sigma_i u/s_n)$. Each term tends to 0 as n increases, provided $s_n \to \infty$, because then we will have $(\sigma_i u/s_n) \to 0$. This means that the series formed by summing the variances σ_i^2 must diverge (and this becomes the first condition). This is not sufficient, however. For example, if we took each σ_i very much greater than the previous ones we could make the ratios σ_n/s_n arbitrarily close to 1 and tending to 1, and the correction term would be $\varepsilon(u)$; this would not be improved by dividing u by s_n. The same problem arises if all the ratios, or an infinite number of them, are greater than some given positive number. To ensure that the correction term tends to 0, it is therefore necessary to have $\sigma_n/s_n \to 0$; this also turns out to be sufficient[45] (and will be the second condition).

45 This is intuitively obvious but it is perhaps best to give the proof, because it is a little less immediate than it might appear at first sight. Fixing $\varepsilon > 0$, we have $\sigma_n/s_n < \varepsilon$ for all n greater than some given N; each σ_i will therefore satisfy $\sigma_i, < \varepsilon s_i < \varepsilon s_n$ for $n > i > N$, and $\sigma_i < s_i \leqslant s_N$ for $i \leqslant N$. Given that $s_n \to \infty$, for all n greater than some given M we have $s_n > s_N/\varepsilon$, that is $s_N/s_n < \varepsilon$, and hence we have, for $i < N$, also $\sigma_i/s_n < s_i/s_n < s_N/s_n < \varepsilon$.

To summarize: *the central limit theorem holds for sums of independent random quantities whose distributions, apart from the variances,*[46] *are the same, provided the total variance diverges* ($s_n \to \infty$) *and the ratios* $\sigma_n/s_n \to 0$. In other words, the theorem holds if, roughly speaking, the contribution of each term becomes negligible compared with that of the total of the preceding terms.

7.7.2. In particular, this holds for bets on Heads and Tails (or at dice, or other games of chance; trials with $p \neq \frac{1}{2}$) when we allow the stakes, S_i to vary from trial to trial. The individual random gains are $S_i(E_i - p)$, the variance is $\sigma_i = S_i \sqrt{(p\tilde{p})}$, and the standardized random quantity is

$$X_i = (E_i - p) / \sqrt{(p\tilde{p})}$$

(for $p = \frac{1}{2}, \sigma_i = \frac{1}{2}S_i$ and $X_i = 2(E_i - \frac{1}{2}) = 2E_i - 1$, as is always used in the case of Heads and Tails).

In order to fix ideas, we can develop considerations of more general validity in the context of this case; this should clarify the result we have obtained. Recall that the σ_i, are the same as the S_i, apart from a change in the unit of measurement.

If the sum of the σ_i^2 were convergent, it would be like having a sum with a finite number of terms (one could stop when the 'remainder' becomes negligible when modifying the distribution obtained). Not only would the argument used to prove that such a distribution is normal not then be valid any longer, but a different argument would even allow one to exclude it being so (except in the trivial case in which all the summands are normal).[47] The condition $s_n \to \infty$ is therefore necessary.

So far as the condition $\sigma_n/s_n \to 0$ is concerned, notice that it is satisfied, in particular, if the σ_n are bounded above (in the above example, this would be the case if the stakes could not exceed some given value) and that this is the only case in which the conclusion holds independently of the order of summation. Were this not the case, one could, in fact, alter the original order, (σ_1, σ_2,..., σ_n,..., in such a way as to every now and again (and hence infinitely often) make the ratio σ_n^2/s_n^2 greater than $\frac{1}{2}$ (say). One possible procedure would be the following: after, say, 100 terms, if the next one (σ_{101}) is too small to give $\sigma_{101}^2/s_{101}^2 > \frac{1}{2}$, insert between the 100th and the 101st the first of the succeeding σ which is $> s_{101} \cdot \sqrt{2}$. Proceed for 100 more terms, and then repeat the operation; and so on.[48]

The conclusion is, therefore, the following: if we have a countable number of summands with no preassigned order, then only the more restrictive condition of *bounded variance* (all the $\sigma_i \leqslant K$) ensures the validity of the central limit theorem (the integers serve as subscripts, but these are merely used by convention to distinguish the summands). On the other hand, if the order has some significance – for example, chronological – then things are different, and the previous conclusion ($s_n \to \infty, \sigma_n/s_n \to 0$) is, in fact, valid and less restrictive.[49]

---

46 See the statement of the theorem for the full meaning of this phrase.

47 By virtue of Cramèr's theorem (Chapter 6, Section 6.12).

48 There is no magic in the figure 100; it was chosen in this example because it seemed best to have a number neither too large, nor too small. The rule must guarantee that all the terms of the original sequence appear in the rearranged sequence (part of the original sequence might be permanently excluded if at each place one term were chosen on the basis of the exigencies of magnitude, or whatever).

49 It seems to be important, both from a conceptual and practical point of view, to distinguish the two cases. In general, however (and, indeed, always, so far as I know), it seems that one only thinks in terms of the case of ordered sequences. It is always necessary to ask oneself whether the symbols actually have a genuine meaning.

If, in particular, we wish to consider the case in which all the stakes (and, therefore, the variances) are increasing, the condition means that the σ_n, must increase more slowly than any geometric progression ('eventually'; i.e. at least from some given point on).

The reason for such bounds is also obviously intuitive. In fact, if a very large bet arises, it, by itself, will influence the shape of the distribution in such a way as to destroy the approach to the normal which might have resulted from all the preceding bets.

It remains to consider now what happens if we let not only the σ_i vary, but also the (standardized) distributions of the X_i. All the expressions that we wrote down in the previous case remain unaltered, except that, in place of $\phi(\sigma_i u)$ and $\varepsilon(\sigma_i u)$, we must now write $\phi(\sigma_i u)$ and $\varepsilon(\sigma_i u)$, allowing for the fact that the distribution (and hence the ϕ and ε) may vary with i.

All that we must do, then, is to examine the 'correction term' in the final expression. The single ε is now replaced by the ε_i and, in order to be able to draw the same conclusion, it will be sufficient to require that the $\varepsilon_i(u)$ all tend to zero in the same way as $\mu \to 0$. In other words, it is sufficient that there exists a positive $\varepsilon(u)$ tending to 0 as $\mu \to 0$, which provides an upper bound for the $\varepsilon_i(u)$; $|\varepsilon_i(u)| \leqslant \varepsilon(u)$.

As far as the meaning of this condition is concerned, it requires that (for the standardized summands, X_i) the masses far away from the origin tend to zero in a sufficiently rapid, uniform manner. More precisely, it requires that $\mathbf{P}(|X_i| \geqslant x)$ be less than some $G(x)$, the same for all the X_i which is decreasing and tending to zero rapidly enough for $\int x^2 |\, \mathrm{d}G(x)| < \infty$ (see Lévy, *Addition*, p. 106).

A sufficient condition is that of Liapounov, which is important from a historical point of view in that it provided the basis of the first rigorous proof of the central limit theorem under fairly unrestrictive conditions (1901). The condition requires that, for at least one exponent $2 + \delta > 2$, the moment exists for all the X_i

$$\mathbf{P}\left(\left|X_i\right|^{2+\delta}\right) = a_i < \infty \tag{7.38}$$

and that

$$\left(a_1 + a_2 + \ldots + a_n\right) / s_n^{2+\delta} \to 0 \quad \text{as} \quad n \to \infty.$$

7.7.3. All that remains is to ask whether the three conditions

$$\left(s_n \to \infty, \sigma_n / s_n \to 0, \left|\varepsilon_i\left(u\right)\right| < \varepsilon\left(u\right)\right)$$

that we know to be sufficient for the validity of the central limit theorem are also necessary. The answer, a somewhat unusual one, perhaps, but one whose sense will become clearer later, is that they are not necessary, but almost necessary.

The necessary and sufficient conditions constitute the so-called Lindeberg–Feller theorem. This improves upon the version which we gave above, and which is known as the Lindeberg–Lévy theorem. The range of questions involved is very extensive and has many aspects, the theory having been developed, more or less independently, and in various ways, by a number of authors, especially in the period 1920–1940. In a certain sense, Lindeberg was the one who began the enterprise, and Lévy and Feller produced the greatest number of contributions (along with Cramèr, Khintchin and many others). Our treatment has looked at just a few of the most important, but straightforward, aspects of the theory. The presentation is original, however, in that

we have made an effort to unify everything (arguments, choice of notation and terminology, emphasis on what is fundamental, and what is peripheral), and because of the inclusion of examples and comments, perhaps novel, but, in any case, probably useful, at least for clarification.

This digression, in addition to providing some historical background, serves to give warning of the impossibility of giving a brief, complete clarification of the phrase, 'not necessary, but almost necessary', which constituted our temporary answer. Not only would we need to include even those things which we intended to omit, but we would also need to give the reasons why we intended omitting them. As an alternative, we shall give the gist of the matter, together with a few examples: the gist is that the conditions only need be weakened a little – 'tinkered with' rather than substantially altered. We have already seen the trivial case (summands all normal) which holds without requiring $s_n \to \infty$; the necessary and sufficient conditions are analogous to the necessary ones but refer to the sums, Y_n, rather than to the summands (allowance is therefore made for intuitive cases of compensation among the effects of different summands, or, for an individual summand, of compensation between a large value of σ_i and a very small $\varepsilon_i(u)$, that is an X_i with a distribution which is almost exactly normal).

An extension of a different kind is provided by the following, which seems, for a variety of reasons, worth mentioning: *a sum of X_h with infinite variances can also tend to a normal distribution* (although within pretty narrow confines, and with rather peculiar forms of normalization). The condition (for X_h with the same distribution F and prevision 0) is that $U(a) = \int_{-a}^{a} x^2 \, \mathrm{d}F$ 'varies slowly' as $a \to \infty$ (that is for every $k > 0$ we must have

$$U(ka)/U(a) \to 1, \quad \text{as} \, a \to \infty,$$

although, by hypothesis, $U(a) \to \infty$). This implies, however, that, *for every* $\alpha < 2$, the moments of order α are finite (this involves the same integral as above, but with $|x|^{\alpha}$ replacing x^2), and that one does not have convergence for the distributions of the $Y_n/\sqrt{n}$ (but instead for some other sequence of constants, to be determined for each case separately). These are the two remarks we made above; note that the second takes up and clarifies the remarks of Chapter 6, 6.7.1, and the first footnote of that section: an example of this is provided by $f(x) = 2|x|^{-3} \log|x| (|x| \geqslant 1)$, where the normalization is given by $Y_n/(\sqrt{n} \log n)$ (see Feller, Vol. II, in several places).

7.7.4. *A complement to the 'law of large numbers'.* This complement (and we present here the important theorem of Khintchin) is included at this point simply for reasons of exposition. In fact, the method of proof is roughly the same as that given above.

We know that for the arithmetic mean, Y_n/n, of the first n random quantities, X_i with $\mathbf{P}(X_i) = 0$, we have $Y_n / n \dot{\to} 0$, and hence $Y_n / n \overset{\leq}{\to} 0$ (the quadratic and weak laws of large numbers, respectively), provided that the variances σ_i^2 are bounded and have a divergent sum. Khintchin's result states that $Y_n / n \to 0$ also holds if the variances are not finite, provided the X_i all have the same distribution. (Other cases also go through, under appropriate restrictions.)

If $\mathbf{P}(X_i) = 0$, we have $\log \phi(u) = u\varepsilon(u)$, with $\varepsilon(u) \to 0$ as $u \to 0$. For Y_n/n, the logarithm of the characteristic function is therefore

$$n \, . \, u/n \, . \, \varepsilon(u/n) = u\varepsilon(u/n) \to 0 \quad \text{as} \, n \to \infty,$$

and hence the characteristic function tends to $e^0 = 1$, and the distribution to $F(x) = (x > 0)$ (all the mass concentrated at the origin): in other words, the limit of Y_n/n (in the weak sense) is 0; $Y_n/n \xrightarrow{<} 0$.

If the distributions of the X_i are not all equal, the $\phi_i(u)$, and therefore the $\varepsilon_i(u)$, will be different. The logarithm of the characteristic function of Y_n/n will then be equal to (u/n) $\sum\varepsilon_i(u/n) = u \times$ the (simple) arithmetic mean of the $\varepsilon_i(u/n)$. Here, too, it suffices to assume that the $\varepsilon_i(u)$ tend to 0 in the same way; that is for all i we must have $|\varepsilon_i(u)| < \varepsilon(u)$, with $\varepsilon(u)$ positive and tending to zero as $u \to 0$. The condition concerning the distributions of the X_i is similar to the previous one (except that it entails the first moment rather than the second): $\mathbf{P}(|X_i| \geqslant x)$ all bounded by one and the same $G(x)$, decreasing and tending to 0 rapidly enough to ensure that $\int x|\mathrm{d}G(x)| < \infty$. Clearly, this is a much less restrictive condition than the previous one: the present condition concerns the influence of the far away masses on the evaluation of the *prevision* (whereas the variance can be infinite); the previous condition was concerned with the influence on the evaluation of the *variance* (which had to exist, or, perhaps better, to be finite).

In order to understand the Khintchin theorem, it is necessary to recollect that we here assume for $\mathbf{P}(X)$ the value given by $\hat{\mathbf{P}}(X)$ (by virtue of the convention we adopted in Chapter 6, 6.5.6).

The theorem states that the mean Y_n/n of $X_1, X_2, \ldots, X_n, \ldots$ (independent, with the same distribution F), converges in a *strong* way to a constant a if and only if the mean value $F(\square) = \hat{\mathbf{P}}(X)$ of F exists and equals a (weak convergence can hold even without this condition).[50]

This property shows $\hat{\mathbf{P}}$ to have an interesting probabilistic significance and hence to appear as something other than merely a useful convention.

50 More general results, with simple proofs, are given in Feller, Vol. II (1966), pp. 231–234.

8

Random Processes with Independent Increments

8.1 Introduction

8.1.1. In Chapter 7, we saw that viewing Heads and Tails as a random process enabled us to present certain problems (laws of large numbers, the central limit theorem) in a more expressive form – as well as giving us insights into their solution. It was, essentially, a question of obtaining a deeper understanding of the problems by looking at them from an appropriate dynamic viewpoint.

This same dynamic viewpoint lends itself, in a natural way, to the study of a number of other problems. Not only does it serve as an aid to one's intuition, but also, and more importantly, it reveals connections between topics and problems that otherwise appear unconnected (a common circumstance, which results in solutions being discovered twice over, and hence not appearing in their true perspective); in so doing, it provides us with a unified overall view.

We have already seen, in the case of Heads and Tails, how the representation as a process enabled us to derive, in an elegant manner, results which could then easily be extended to more general cases. We now proceed by following up this idea in two different, but related, directions:

making precise the kind of process to be considered as a first development of the case of Heads and Tails;
considering (first for the case of Heads and Tails, and then in the wider ambit mentioned above) problems of a more complicated nature than those studied so far.

8.1.2. The random processes that we shall consider first are those *with independent increments*,[1] and we shall pay special attention to the *homogeneous* processes. This will be the case both for processes *in discrete time* (to which we have restricted ourselves so far) and for those in *continuous time*.

In *discrete time* (where t assumes integer values), the processes will be of the same form as those already considered: $Y(t) = X_1 + X_2 + \ldots + X_t$, the sum of *independent* random quantities (increments) X_i. In terms of the Y, one can describe the process by

1 Here, and in what follows, *independent* always means *stochastically independent*.

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

saying that the increments in $Y(t)$ over disjoint time intervals are independent: that is $Y(t_2) - Y(t_1)$ and $Y(t_4) - Y(t_3)$ are independent for $t_1 < t_2 \leqslant t_3 < t_4$. The increments are, in fact, either the Xs themselves, or sums of distinct Xs, according to whether we are dealing with unit time, or with a longer time interval.

Such a process is called *homogeneous* if all the X_i have precisely the same distribution, F. More generally, all increments $Y(t + \tau) - Y(t)$ relating to intervals with the same length, τ, have exactly the same distribution (given by the convolution $F^{\tau *}$).

In *continuous time* (a case which we have so far only mentioned in passing) the above conditions, expressed in terms of the increments of $Y(t)$, remain unchanged; except that now, of course, they must hold for arbitrary times t (instead of only for integer values, as in the discrete time case). This makes the conditions *much more restrictive*. The consequence of this is that whereas in the discrete case all distributions were possible for the Xs (and all distributions decomposable into a t-fold convolution were admissible for the $Y(t)$), in continuous time we can only have, for the $Y(t)$ and their increments $Y(t + \tau) - Y(t)$, those distributions whose decomposition can be taken a great deal further: in other words, the *infinitely divisible* distributions (which we mentioned in Chapter 6, 6.11.3).[2]

In such a process, the function $Y(t)$ can be thought of as decomposed into two components, $Y(t) = Y_J(t) + Y_C(t)$, the first of which varies *by jumps*, and the second *continuously*. The arguments we shall use, based on this idea, incomplete though they are at the moment, are essentially correct so far as the conclusions are concerned, although critical comments are required at one point (Section 8.3.1) in connection with the interpretation of these conclusions and the initial concepts.

The conclusion will be that the distribution of $Y(t)$ (or of the increment $Y(t + \tau) - Y(t)$, respectively) can be completely and meaningfully expressed by considering the two components Y_J and Y_C separately:

in order to characterize the distribution of $Y_J(t)$ it is necessary and sufficient to give *the prevision*, over the interval $[0, t]$ ($[t, t + \tau]$, respectively), *of the number of jumps of various sizes*;
in order to characterize the distribution of $Y_C(t)$ it is necessary and sufficient to give *the prevision and the variance of the continuous component* (since, as we shall see, its distribution is necessarily *normal*);
taken together, these characterize $Y(t)$.

All the previsions are additive functions of the intervals and, except for that of the increment of the continuous component, Y_C, they are essentially non-negative, and therefore nondecreasing.

In the *homogeneous* case, they depend only on the length of the interval, and are therefore proportional to it.

8.1.3. We shall illustrate straightaway, by presenting some of the most important cases, the structure that derives from what we have described as the most general form

2 These statements are not quite correct as they stand. Firstly, X_i could be a *certain* number; secondly, the jumps might occur at *known* instants. These are trivial special cases, however, which could be considered separately (see Section 8.1.4).

of random process with independent increments. Although this will only be a summary, it should help to make clear the scope of our investigation and also give an idea of the kinds of problems we shall encounter.

Let us first restrict ourselves to the *homogeneous* case (it will be seen subsequently (Section 8.1.4) that the extension to the general case is reasonably straightforward and only involves minor additional considerations).

It is convenient to indicate and collect together at this point the notation that will be used in what follows. This will be given for the homogeneous case; hence a single distribution suffices for the increments $Y(t_0 + t) - Y(t_0)$. This will depend on t but not on t_0 and will also be the distribution of $Y(t)$ if the initial condition $Y(0) = 0$ is assumed (as will usually be convenient). The distribution function, density (if any) and characteristic function of this distribution will be denoted by $F^t(y)$, $f^t(y)$ and $\phi^t(u)$ respectively. The t is, in fact, an exponent of $\phi(u)$ (as is obvious from the homogeneity and the independence of the increments) and is used as a superscript for F and f both for uniformity and to leave room for possible subscripts. Its use is also partially justified by the fact that F^t and f^t are, actually t-fold convolutions, $(F^1)^{t*}$ and $(f^1)^{t*}$, of F^1 and f^1 with themselves, provided that the concept (where it makes sense, as is the case here for all t, because of infinite divisibility) is suitably extended to the case of noninteger exponent t. The distribution and density (if any) of the *jumps* will be denoted by $F(x)$ and $f(x)$, respectively (with no superscripts), and the characteristic function by $\chi(u)$.

Let us examine now the various cases.

The simplest example, and the one which forms the basis for the construction of all the random processes of the type under consideration, is that of the *Poisson process*. This is a jump process, all jumps being of the same size, x; for the time being, we shall take $x = 1$, so that $Y(t) = N(t) =$ the *number of jumps* in $[0, t]$ (see Figure 8.1). We shall denote by μ the *prevision* of the number of jumps occurring per unit time (i.e. μt is the prevision of the number of jumps in a time interval of length t). In an infinitesimal time, dt, the prevision of the number of jumps is μdt; up to an infinitesimal quantity of greater order, this is also the probability that *one* jump occurs within the small time interval (the probability of more than one jump occurring is, in comparison, negligible). We call μ the *intensity* of the process.

The distribution of $N(t)$, the number of jumps occurring before time t, is Poisson, with prevision $a = \mu t$ (see Chapter 6, 6.11.2, equation 6.39, for the explicit form of the probabilities and the characteristic function).

We recall that the variance in this case is also equal to μt. It is better, however, to note explicitly that the prevision is $x\mu t$, and the variance is $x^2\mu t$, where x is the magnitude of the jump. In this way, one avoids the ambiguities which arise from ignoring dimensional questions (i.e. taking x = pure number) and, subsequently, from assuming the special value $x = 1$ (for which $x = x^2$).

A superposition of several Poisson processes, having jumps of different sizes, x_k, and different intensities, μ_k ($k = 1, 2, \ldots, m$), is also a jump process, homogeneous with independent increments, $Y(t) = \Sigma x_k N_k(t)$. It also has an alternative interpretation as a process of intensity $\mu = \mu_1 + \mu_2 + \ldots + \mu_m$, where each jump has a random size X (independently of the others), X taking the value x_k with probability μ_k/μ.

Instead of considering the sum of a finite number of terms, we could also consider an infinite series, or even an integral (in general, a Stieltjes integral). Provided the total intensity remains finite, the above interpretations continue to apply (except that X now

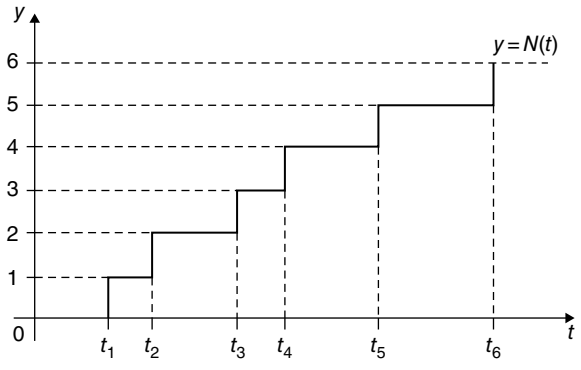


Figure 8.1 The simple Poisson process.

has an arbitrary distribution, rather than the discrete one given above). This case is referred to as a *compound Poisson process* and provides the most general homogeneous process with independent increments and *discrete jumps* (i.e. finite in number in any bounded interval).

Using the same procedure, one can also obtain the *generalized Poisson processes*: that is those *with everywhere dense jumps* (discontinuity points), with $\mu = \infty$. It is necessary, of course, to check that the process does not diverge: for this, we require that the intensity μ_ε of jumps X with $|X| > \varepsilon (\varepsilon > 0)$ remains finite, diverging only as $\varepsilon \to 0$, and then not too rapidly.

The Poisson processes, together with the compound and generalized forms, exhaust all the possibilities for the jump component $Y_J(t)$.

It remains to consider the continuous component $Y_C(t)$. In every case, $Y(t)$ could be considered as the sum of N increments (N arbitrary), corresponding to the N small intervals of length t/N into which the interval $[0, t]$ could be decomposed. If one makes precise the idea of separating off the 'large' increments (large in absolute value), which correspond to the jumps, it can be shown that the others (the 'small' increments) satisfy the conditions of the central limit theorem, and hence the distribution is necessarily normal (as we mentioned above).

As we have already stated, its prevision varies in a linear fashion, $\mathbf{P}[Y_C(t)] = mt$, and the same is true of the variance, $\mathbf{P}[\{Y_C(t) - mt\}^2] = \sigma^2 t$ (where m and σ^2 denote the prevision and variance corresponding to $t = 1$).[3] In some cases, it may also turn out to be convenient to separate the certain linear function, mt, and the fair component (with zero prevision), $Y_C(t) - mt$ (whose variance is still $\sigma^2 t$).

The set-up we have just described is called the *Wiener–Lévy process* (see also Chapter 7, 7.6.5).

8.1.4. So far as the increment over some given interval is concerned, the conclusions arrived at in the homogeneous case carry over without modification to the general case. This is clear, because the conclusions depend on the prevision (of the number of certain jumps etc.) over such an interval as a whole, and not on the way the prevision is distributed

3 These formulae hold, of course, for every $Y(t)$ that is homogeneous with independent increments (with finite m and σ^2). We mention them here, for the particular case of Y_C, only because of their importance in specifying the distribution in this case.

over the subintervals. In the inhomogeneous case, the only new feature is that, within an interval, each prevision can be increasing in an arbitrary way, not necessarily linearly.

The one thing that is required is that one exclude (or, better, consider separately, if they exist) the points of discontinuity for such a prevision. In fact, at such points one would have a nonzero probability of discontinuity (a 'fixed discontinuity'; see the remark in the footnote to Section 8.1.2). This is equivalent to saying that at these points $Y(t)$ receives a random (instantaneous) increment, incompatible with the nature of the 'process in continuous time', because (being instantaneous) it is not decomposable, and therefore does not have to obey the 'infinite divisibility' requirement. In what follows, we shall always tacitly assume that such fixed discontinuities have been excluded.

Observe that, if all the previsions are proportional to one another, the process can be said to be homogeneous with respect to a different time scale (one which is proportional to them). In general, however, the previsions will vary in different ways, and then we have no recourse to anything of this kind.

8.1.5. Let us now consider more specifically some of the problems that one encounters. These are of interest for a number of reasons : by virtue of their probabilistic significance and their range of application; because of the various mathematical aspects involved; and, above all, because of the unified and intuitively meaningful presentation that can be given of a vast collection of seemingly distinct problems. The problems that we shall mention are only a tiny sample from this collection and the treatment that we shall give will only touch upon some of the more essential topics, presenting them in their simplest forms.

First of all, we have to translate into actual mathematical terms (through the distribution, by means of the characteristic function) the characterization of the general process with independent increments, and hence of the most general infinitely divisible distribution (which we have already given above, in terms of the intensities of the jumps and the normal component).

Particular attention must be paid, however, to the limiting arguments that lead to the generalized Poisson process, since the latter gives rise to rather different problems. For example, in our preliminary remarks we did not mention that sometimes convergence can only be achieved by 'compensating' the jumps by means of a certain linear function and that, in this case, the intuitive idea of behaviour similar to that of the discrete case (apart from minor details) must undergo a radical change.

Even the behaviour of the continuous component (the Wiener–Lévy process), despite what one might at first sight be led to expect from the regular and familiar shape of the normal distribution, turns out to be extremely 'pathological'. The study of the behaviour of the function (or, more precisely, the behaviour which it 'almost certainly' enjoys) is, however, a more advanced problem. We shall begin by saying something about the distribution.

How does the distribution of $Y(t)$ behave as t increases? We already know that it tends to normality in the case of finite variances, but there also exist processes (of the generalized Poisson type) with infinite variances. The answer in this case in that there exist other types of stable distribution, all corresponding to generalized Poisson processes (more precisely, as we shall see in Sections 8.4.1–8.4.4, there are a doubly infinite collection of them, reducible, essentially, to a single infinite collection). The processes which are not stable either tend to a stable form, or do not tend to anything at all.

The key to the whole question lies in the behaviour at the other extreme, as $t \to 0$. This is directly connected with the intensity of the jumps of different sizes: one has stability if the intensity of the jumps $>x$ or $<-x$ decreases proportionally to $x^{-\alpha}(x > 0, 0 < \alpha < 2)$; one has a tendency to a stable form if the process, in some sense, approximately satisfies this condition. Referring back to a remark made previously, we note that the 'sum of the jumps' does not require 'compensation' if $\alpha < 1$, but does if $\alpha > 1$ ($\alpha = 1$ constitutes a subcase of its own). It follows that the stable distributions generated by just the positive (negative) jumps extend only over the positive (negative) values if $\alpha < 1$, whereas they extend over all positive and negative values if $\alpha \geqslant 1$.

An important special case is that of 'lattice' processes, in which $Y(t)$ can only assume integer values (or those of an arithmetic progression; there is no essential difference). They must, of course, be compound Poisson processes, but this does not mean that they cannot tend to a stable (continuous distribution as t increases. Indeed, if the variances are finite they necessarily tend to the normal distribution (as we already saw, in the first instance, in the case of Heads and Tails).

8.1.6. The study of the way in which the distribution of $Y(t)$ varies with t does not necessarily entail the study of the behaviour of the actual process $Y(t)$ (i.e. of the function $Y(t)$). The essential thing is to examine the characteristics of the behaviour of the latter, which are of interest to us, behaviour which can only be studied by the simultaneous consideration and comparison of values of $Y(t)$ at different times t (possibly a large or infinite number of them).

So far as the phrase 'infinite number' is concerned, we should make it clear at once that it means 'an arbitrarily large, but finite, number' (unless, in the case under consideration, one makes some additional assumption, such as the validity of countable additivity, or gives some further explanation). However, in order to avoid making heavy weather of the presentation with subtle, critical arguments, we shall often resort to intuitive explanations of this kind (as well as to the corresponding 'practical' justifications).

Problems that have already been encountered in the discrete cases – like that of the asymptotic behaviour of $Y(t)/t$ (as $t \to \infty$), to which the strong law of large numbers gives the solution – are restated and shown to have, generally speaking, similar solutions in the continuous case. One also meets, in the latter case, problems that are, in a certain sense, reciprocal; involving what happens as $t \to 0$ ('local' behaviour – like 'continuity' etc. – at the origin and, hence, in the homogeneous case, at any arbitrary point in time). In the Wiener–Lévy process, the two problems correspond exactly to one another by reciprocity.

If one confines attention to considerations based on the Tchebychev inequality, the conclusions hold for every homogeneous process with independent increments for which the variance is finite. One such process, viewing the jumps and possible horizontal segments 'in the large' (i.e. with respect to large time intervals and intervals of the ordinate), in such a way as to make them imperceptible, is the Wiener–Lévy process. Indeed, Lévy calls it a Brownian motion process, which corresponds to the perception of an observer who is not able to single out the numerous, tiny collisions that, in any imperceptibly small time interval, suddenly change the motion of the particle under observation.

More generally, even for arbitrary random processes (provided they have finite variance), answers can be obtained to a number of problems, even though, in general, they may only be qualitative and based on second-order characteristics (as in the case of a

single random quantity, or of several). Here, the random quantities to be considered are the values $Y(t)$ and the characteristics to be used are the previsions, variances and covariances (or, equivalently, $\mathbf{P}[Y(t)]$, $\mathbf{P}[Y^2(t)$, $\mathbf{P}[Y(t_1)Y(t_2)]$). In our case, this is trivial: evaluated at $t = 0$, the prevision and variance of $Y(t)$ are, as we know, mt and $\sigma^2 t$, and the covariance of $Y(t_1)$ and $Y(t_2)$ is $\sigma^2 t_1$ (if $t_1 \leqslant t_2$); the correlation coefficient is, therefore, $r(t_1, t_2) = \sigma^2 t_1/\sigma\sqrt{t_1}\cdot\sigma\sqrt{t_2} = \sqrt{(t_1/t_2)}$ (it is sufficient to observe that $Y(t_2) = Y(t_1) + [Y(t_2) - Y(t_1)]$, and that the two summands are independent).

There are other problems, however, that require one to take all the characteristics into account and to have recourse to new methods of approach. (This also applies to the problems considered previously, if one wishes to obtain conclusions that are more precise, in a quantitative sense, and more specifically related to the particular process under consideration.) A concept that can usefully be applied to a number of problems is that of a *barrier* (a line in the (t, y)-plane on which $y = Y(t)$ is represented). One observes when the barrier is first reached, or when it is subsequently reached, or, sometimes, one assumes that the barrier modifies the process (it may be absorbing – i.e. the process stops – or reflecting, and so on).

The classical problem of this kind, and one that finds immediate application, is that of the 'gambler's ruin' (corresponding to the point when his gain reaches the level $-c$, where c is his initial capital). There are a number of obvious variants: one could consider the capital as being variable (an arbitrary barrier rather than the horizontal line $y = -c$), or one could think of two gamblers, both having a bounded initial capital (or variable capital), and so on.

In addition to this and other interesting and practical applications, problems of this kind also find application in studying various aspects of the behaviour of the function $Y(t)$. In particular, and this question has been studied more than any other, they are useful for specifying the asymptotic behaviour, indicating which functions tend to zero too rapidly, or not rapidly enough, to provide a (practically certain) bound for $Y(t)/t$ from some point $t = T$ on (and similarly as $t \to 0$).

Reflecting barriers, and others which modify the process, take us beyond the scope of this present chapter. In any case, they will enter, in a certain sense, into the considerations we shall make later, and will provide an instructive and useful technique. In particular, they will enable us to make use of the elegant and powerful arguments of Desiré André (and others) based on symmetries.

8.2 The General Case: The Case of Asymptotic Normality

8.2.1. Let us first give a precise, analytic statement of what we have hitherto presented in a descriptive form concerning the structure and properties of the general homogeneous process with independent increments.

When the intensity of the jumps, μ, and the variance, σ^2, are finite, the process is either normal (if $\mu = 0$), or asymptotically normal (as we have already seen from the central limit theorem). In other words, we either have the Wiener–Lévy process, or something which approximates to it asymptotically. We are not saying that the restrictions made are necessary for such behaviour: the restriction on μ, that is $\mu < \infty$, has no direct relevance, and the restriction $\sigma^2 < \infty$ could be weakened somewhat (see Chapter 7, 7.7.3). However, for our immediate purpose it is more appropriate to

concentrate attention on the simplest case, avoiding tiresome complications which do not really contribute anything to our understanding.

Our first task, once we have set up the general analytical framework, is to provide insight into the way in which, and in what sense, processes of this kind (i.e. those which are asymptotically normal) can be considered as an approximation, on some suitable scale, to the Wiener–Lévy process, and conversely. In this way, any conclusion established for a special case, for example, that of Heads and Tails, turns out to be necessarily valid (in the appropriate asymptotic version) in the general case. Note that this enables us, among other things, to establish properties of the Wiener–Lévy process by means of elementary combinatorial techniques, which are, in themselves, applicable only to the case of Heads and Tails. Conversely, it enables us to derive properties of the latter, and of similar examples, that cannot be obtained directly (for example, asymptotic properties) by using approaches – often much simpler – which deal with the limit case of the Wiener–Lévy process (itself based on the normal distribution). This is just one example, out of many, where the possibility exists of advantageously switching from a discrete schematization to a continuous one, or vice versa, as the case may be.

8.2.2. The *Wiener–Lévy process* can be derived as the limit case of the *Heads and Tails* process (in discrete time).

Suppose, in fact, that we change the scale, performing tosses at shorter time intervals, and with smaller stakes, so that the variance per unit time remains the same. For this to be so, if the stake is reduced by a factor of N ($a = 1/N$) the number of tosses per unit time must be increased to N^2 (with time intervals $\tau = 1/N^2$). To see this, note that the variance for each small time interval τ is $a^2 = (1/N)^2$, and in order for it to be equal to 1 per unit time, the number of small intervals by which it has to be multiplied must be N^2.

By taking N sufficiently large, one can arrange things in such a way that the increment per unit time has a distribution arbitrarily close to that of the standardized normal. Taking N even larger, one can arrange for the same properties to hold, also, for the small intervals. In other words, one can arrange for the distribution of $Y(t)/\sqrt{t}$ to be arbitrarily close, to any preassigned degree, to the standardized normal, for every t exceeding some arbitrarily chosen value. This could also be expressed by saying that the Heads and Tails process can be made to resemble (with a suitable change of scale) a process in discrete time with normally distributed jumps, and (with a more pronounced change of scale) even to resemble the Wiener–Lévy process (provided, of course, that one regards as meaningless any claim that the scheme is valid, or, anyway, observable, for arbitrarily small time periods).

In terms of the characteristic function, these considerations reduce, in the first case, to the straightforward and obvious observation that if we substitute e^{-u^2} for $\cos u$ then $[\cos(u/\sqrt{n}]^n \to e^{-u^2}$ becomes the identity $\left[e^{-(u/\sqrt{n})^2}\right]^n = e^{-u^2}$; whereas, in the second case, they simply repeat the procedure used in Chapter 7, 7.7.1. In the Heads and Tails process, $Y(t)$ has characteristic function $\phi^t(u) = (\cos u)^t$ (t = integer); under the above-mentioned change of scale, this becomes $\left[\cos(u/N)\right]^{tN^2}$ (t = integer/N^2) and in the limit (as $N \to \infty$) it becomes e^{-u^2} (t arbitrary).

8.2.3. The Wiener–Lévy process can be obtained in an analogous manner from the *Poisson version of the Heads and Tails process* (a compound Poisson process, with the intensity of the jumps given by $\mu = 1$, and with jumps ± 1, each with probability $\frac{1}{2}$).

The difference is that instead of there certainly being a toss after each unit time interval the tosses occur at random, with a *prevision* of one per unit time (the probability being given by dt in each infinitesimal time interval of length dt). Alternatively (as we mentioned already in Section 8.1.3), we can say that $Y(t) = Y_1(t) - Y_2(t)$, where Y_1 and Y_2 are the number of positive and negative gains, respectively, both occurring at random and independently, each with intensity $\frac{1}{2}$.

The distribution of $Y(t)$ in such a process is the Poisson mixture of the distributions of Heads and Tails. In terms of the characteristic function, $\phi^t(u)$ is the Poisson mixture (with 'weights' given by the probabilities $e^{-t}t^n/n!$ of there being n jumps in $[0, t]$) of the $(\cos u)^n$ (the characteristic functions of the sums of n jumps; i.e. of $Y(t)$, assuming that there are n jumps up until time t):

$$\phi^t(u) = \sum e^{-t}(t\cos u)^n/n! = e^{t(\cos u - 1)}. \tag{8.1}$$

In this case, too, the same change of scale (jumps reduced by $1/N$, and the intensity increased by N^2) leads to the Wiener–Lévy process. In fact, as $N \to \infty$, the characteristic function $\exp\{tN^2[\cos(u/N) - 1]\}$ tends to e^{-tu^2}.

We therefore obtain the conclusion mentioned above, and it is worth pausing to consider what it actually means. It establishes that the distribution of the gain in a game of Heads and Tails is, after a sufficiently long period of time has elapsed, practically the same, no matter whether tosses were performed in a regular fashion (one after each unit time period), or were randomly distributed (with a Poisson distribution, yielding, *in prevision*, one toss per each unit time period).

8.2.4. The three examples given above (the Heads and Tails process, in both the regular, discrete case and its Poisson variant, and the normally distributed jump process in discrete time) provide the simplest approaches to approximate representations of the Wiener–Lévy process and should be borne in mind in this connection. If we wished, we could also include a fourth such example; the Poisson variant of the normal jump process:

$$\phi^t(u) = \sum e^{-t}\left(t\,e^{-u^2}\right)^n/n! = \exp\left[t\left(e^{-u^2} - 1\right)\right]. \tag{8.2}$$

Strictly speaking, however, if one ignores the psychological case for presenting these introductory examples, the above discussion is entirely superfluous. We have merely anticipated, in a few special cases, ideas which can be examined with equal facility in the general case.

Let us now turn, therefore, to a systematic study of the general case. We begin with the Poisson process, and then proceed to a study of the compound Poisson processes.

8.2.5. The (simple) *Poisson process* deals with the number of occurrences, $N(t)$, of some given phenomenon within a time period $[0, t]$. In other words, it counts the *jumps*, each considered as being of unit size (like a meter that clicks once each time it records a phenomenon – such as the beginning of a telephone conversation, a particle hitting a screen, a visitor entering a museum, or a traveller entering an underground station etc.).

The conditions given in Section 8.1.3 simply mean that we must be dealing with a homogeneous process with independent increments, and with jumps all of unit size.

It is instructive to go back to these conditions and, in the context of the present problem of deriving the probabilities of the Poisson distribution, to provide two alternative derivations in addition to that given in Chapter 6, 6.11.2.

First method. Let a ($a = \mu$) be the prevision of the number of jumps occurring in a given interval (of length t, where μ denotes the intensity). If we subdivide the interval into n equal parts (n large, so that a/n is small compared with 1; i.e. $a/n < \varepsilon$, for some preassigned $\varepsilon > 0$), then a/n is the prevision of the number of jumps occurring in each small interval. We also have $(a/n) = q_n m_n$, where q_n is the probability of *at least one* jump occurring in a small time interval of length t/n, and m_n is the prevision of the number of jumps occurring in intervals containing at least one jump. It follows that we must have $m_n \to 1$ (as $n \to \infty$). If this were not so, it would mean that each discontinuity point had a positive probability of having further jumps in any arbitrarily small neighbourhood of itself; in other words, practically speaking, of being a multiple jump (contrary to the hypothesis that all jumps are of unit size).[4]

The probability that h out of the n small intervals contain discontinuities, and $n - h$ do not, is given by $\binom{n}{h} q_n^h \left(1 - q_n\right)^{n-h}$. As $n \to \infty$, the probability that there are small intervals containing more than one jump becomes negligible (so that h gives the actual number of jumps). On the other hand, we also have $q_n \simeq a/n$ and, therefore

$$p_h(t) = \lim \binom{n}{h} \left(\frac{a}{n}\right)^h \left(1 - \frac{a}{n}\right)^{n-h}.$$

In this way, we reduce to the formulation and procedure that we have already seen (in Chapter 6, 6.11.2) for so-called 'rare events' (which the occurrences of jumps in very small intervals certainly are).

Second method. We can establish immediately that $p_0(t)$, the probability of no jumps in a time interval of length t, must be of the form e^{-kt}. To see this, we note that, because of the independence assumption,

$$p_0(t' + t'') = p_0(t')\, p_0(t'')$$

and this relation characterizes the exponential function. The probability that the *waiting time*, T_1, until the occurrence of the first jump does not exceed t is given by $F(t) = 1 - p_0(t) = 1 - e^{-kt}$ (which is equivalent to saying that it is not the case that no jump occurred between 0 and t). From the distribution function $F(t)$, we can derive the density function $f(t) = k\, e^{-kt}$, and we then know that the characteristic function is given by

$$\phi(u) = 1/(1 - ku).$$

4 It would certainly be more direct to impose an additional condition requiring that the probability $p^*(t) = 1 - p_0(t) - p_1(t)$ of there being *two or more* jumps in an interval of length t be an infinitesimal of second order or above. This is, in fact, the approach adopted in many treatments, but it carries the risk of being interpreted as an additional restriction, without which there could be different processes, each compatible with the initial assumptions.

We recall (although, in fact, it follows directly from the above) that the gamma distribution is obtained by convolution:

$$\left[\phi(u)\right]^h = (1-ku)^{-h}, \quad f^{h^*} = Kx^{h-1}e^{-kt}\ (x \geqslant 0),$$
$$F^{h^*} = 1 - e^{-kt}\left[1 + \frac{kt}{1!} + \frac{(kt)^2}{2!} + \ldots + \frac{(kt)^h}{h!}\right].$$

This, therefore, gives the distribution of S_h, the waiting time for the occurrence of the hth jump, which is the sum of the first h waiting times (independent, and exponentially distributed):

$$S_h = T_1 + T_2 + \ldots + T_h.$$

This method of approach is, in a sense, the converse of the first one. The connection is provided by noting that $N(t) < h$ is equivalent to $S_h > t$ ('less than h jumps in $[0, t]$' = 'the hth jump takes place after time t'), and that $N(t) = h$ is equivalent to

$$\{N(t) < h+1\} - \{N(t) < h\} = \{S_{h+1} > t\} - \{S_h > t\}.$$

From this we see that

$$p_h(t) = \mathbf{P}\{N(t) = h\} = \mathbf{P}\{S_{h+1} > t\} - \mathbf{P}\{S_h > t\} = e^{-kt}(kt)^h/h!, \tag{8.3}$$

again yielding the Poisson distribution. Given that its prevision is kt, it turns out that k, introduced as an arbitrary constant, is actually μ: imagine the latter in place of k, therefore, in the preceding formulae.

Third method. This is perhaps the most intuitive approach and the most useful in that it can be applied to any scheme involving passages through different 'states' with intensities μ_{ij}, constant or variable, where $\mu_{ij}\,dt$ = the probability that from an initial state 'i' at time t one passes to state 'j' within an infinitesimal time dt.

Let us denote by $p_h(t)$ the functions (assumed to be unknown) that express the probabilities of being in state h at time t (in the Poisson process, state h at time t corresponds to $N(t) = h$, $h = 0, 1, 2, \ldots$). In the general case, the probability of a passage from i to j in an infinitesimal time dt is given by $p_i(t)\mu_{ij}\,dt$ (the probability of two passages, from i to some h, and then from h to j, within time dt, is negligible, since it is an infinitesimal of the second order). The change in $p_h(t)$ is given by $dp_h = p_h'\,dt$ and consists of the positive contribution of all the incoming terms (from all the other i to h), and the negative contribution of the outgoing terms (from h to all the other j). One has, therefore, in the general case (which has been mentioned merely to provide a proper setting for the case of special interest to us), a system of differential equations (which requires the addition of suitable initial conditions).

Our case is much simpler, however: we have only one probability, that of passing from an h to the next $h + 1$, the intensity remaining constant throughout. F or $h = 0$, we have only the outgoing term, $-\mu p_0(t)\,dt$, whereas for $h > 0$, we have, in addition to $-\mu p_h(t)\,dt$, the incoming term $\mu p_{h-1}(t)\,dt$. This reduces to the (recursive) system of equations

$$p_0'(t) = -\mu p_0(t), \qquad p_h'(t) = \mu p_{h-1}(t) - \mu p_h(t), \tag{8.4}$$

together with the initial conditions, $p_0(0) = 1$, $p_h(0) = 0$ $(h \neq 0)$.

From the first equation, we obtain immediately that $p_0(t) = e^{-\mu t}$ and, hence, from the second equation we obtain

$$p_1(t) = \mu t \ e^{-\mu t},$$

and so on. If (realizing from the first terms that it is convenient to extract the factor $e^{-\mu t}$) we set $p_h(t) = e^{-\mu t} g_h(t)$ (with $g_0(t) = 1$, and $g_h(0) = 0$ for $h \neq 0$), we can virtually eliminate any need for calculations: the recursive relation for the $g_h(t)$ reduces to the extremely simple form $g_h'(t) = \mu g_{h-1}(t)$, so that $g_h(t) = (\mu t)^h / h!$.

8.2.6. As an alternative to the method used in Chapter 6, 6.11.2, the characteristic function of the Poisson distribution can be obtained by a direct calculation:

$$\phi^t(u) = \sum_h e^{-\mu t} (\mu t)^h e^{iuh} / h! = \exp\left\{\mu t\left(e^{iu} - 1\right)\right\}. \tag{8.5}$$

So far as the random process is concerned, it is very instructive and meaningful to observe that, as $t \to 0$, we have, asymptotically,

$$\phi^t(u) = 1 + \mu t\left(e^{iu} - 1\right) = (1 - \mu t) + \mu t \ e^{iu}$$

(probability $1 - \mu t$ of 0, and μt of 1). This is the 'infinitesimal transformation' from which the process derives. The simplest way of seeing this is, perhaps, to observe that

$$\phi^t(u) = \lim_{n\to\infty}\left[1 + \frac{1}{n}\mu t\left(e^{iu} - 1\right)\right]^n.$$

The Poisson process also tends to a normal form (as is obvious, given that it has finite variance); in other words, asymptotically it approximates the Wiener–Lévy process. However, the prevision is no longer zero, but equals μt (as does the variance). In order to obtain zero prevision, that is in order to have a finite process, it is necessary to subtract off a linear term and to consider a new process consisting of $N(t) - \mu t$: instead of the number of jumps, one considers the difference between this number and its prevision. The behaviour of

$$Y(t) = N(t) - \mu t$$

gives rise to the saw-tooth appearance of Figure 8.4:[5] all the jumps are equal to +1, and the segments in between have slope $-\mu$. With the introduction of the correction term $-\mu t$, the characteristic function is multiplied by $e^{-i\mu t u}$, and we obtain

$$\exp\left\{\mu t\left(e^{iu} - 1 - iu\right)\right\}. \tag{8.6}$$

This was obvious: when we take the logarithm of the characteristic function, the linear term in u must vanish, and we have

$$e^{iu} - 1 - iu = -\tfrac{1}{2}u^2\left[1 + \varepsilon(u)\right] \qquad \left(\varepsilon(u) \to 0 \qquad \text{as } u \to 0\right).$$

5 Figure 8.4 has been placed later in the text (Section 8.4.3), in order to emphasize its connection with Figure 8.5.

Replacing u by $u/\sqrt{(\mu t)}$, in order to obtain the standardized distribution, we obtain

$$\exp\left\{\mu t\left[-\tfrac{1}{2}\left(u/\sqrt{(\mu t)}\right)^2\right]\left[1+\varepsilon\left(u/\sqrt{(\mu t)}\right)\right]\right\}\to \mathrm{e}^{-\frac{1}{2}u^2}.$$

This provides, if required, a fifth approach to approximating the Wiener–Lévy process. Simple though it is, we thought it worth spending some time on this example, because the idea of adjusting (in the mean) the jumps by a certain linear term, that is of considering jumps with respect to an inclined line rather than a horizontal one, turns out, in a number of cases, to be necessary in order to ensure the convergence of certain procedures (and we shall see examples of this in Sections 8.2.7 and 8.2.9).

8.2.7. The compound Poisson process can be developed along the same lines as were followed in the case of the (simple) Poisson process. The analysis of the most general cases can then be attempted, recognizing that these are, in fact, the generalized Poisson processes.

In the case of a compound Poisson process – with intensity μ, and each jump X having distribution function $F(x)$ and characteristic function $\chi(u) = \mathbf{P}(\mathrm{e}^{\mathrm{i}uX})$ – the characteristic function $\phi^t(u)$ is obtained in exactly the same way as for the simple Poisson process: that is by substituting $\chi(u)$ in place of $\mathrm{e}^{\mathrm{i}u}$ (the latter being the $\chi(u)$ of the simple case, where $X = 1$ with certainty; that is $F(x)$ consists of a single mass concentrated at the point $x = 1$).

This is immediate: one can either note that the 'infinitesimal transformation' is now $1 + \mu t[x(u) - 1] = (1 - \mu t) + \mu t\chi(u)$ (probability $1 - \mu t$ of 0, and probability μt distributed according to the distribution of a jump), from which it follows that

$$\phi^t(u)=\lim_{n\to\infty}\left\{1+\frac{1}{n}\mu t\left[\chi(u)-1\right]\right\}^n=\exp\left\{\mu t\left[\chi(u)-1\right]\right\}, \tag{8.7}$$

or one can simply observe that, conditional on the number of jumps $N(t)$ being equal to h, the characteristic function is given by $\chi^h(u)$, and hence that $\phi^t(u)$ is a mixture of the latter, with weights equal to the probabilities of the individual h. In other words,

$$\phi^t(u)=\sum_{h=0}^{\infty}\left[e^{-\mu t}(\mu t)^h/h!\right]\chi^h(u)=\mathrm{e}^{-\mu t}\sum_{h=0}^{\infty}\left[\mu t\chi(u)\right]^h/h!, \tag{8.8}$$

which is the series expansion of the form given in (8.7). Here too, of course, we are merely rewriting (8.5) with $\chi(u)$ substituted in place of $\mathrm{e}^{\mathrm{i}u}$.

If one wishes to give an expression in terms of the distribution of the jumps, $F(x)$, one can write the characteristic function in the form

$$\phi^t(u)=\exp\left\{\mu t\left[\int \mathrm{e}^{\mathrm{i}ux}\,\mathrm{d}F(x)-1\right]\right\}=\exp\left\{t\int\left(\mathrm{e}^{\mathrm{i}ux}-1\right)\mu\,\mathrm{d}F(x)\right\}, \tag{8.9}$$

or, alternatively,

$$\phi^t(u)=\exp\left\{t\int\left(\mathrm{e}^{\mathrm{i}ux}-1\right)\mathrm{d}M(x)\right\}, \tag{8.10}$$

where $M(x) = \mu F(x)$. As another alternative (in order to deal more satisfactorily with the arbitrariness of the additive constant), we could take

$$\begin{aligned} M(x) &= \mu F(x) && \text{for } x < 0, \\ M(x) &= \mu\big[F(x) - 1\big] = -\mu\big[1 - F(x)\big] && \text{for } x > 0, \end{aligned} \tag{8.11}$$

so that, to make the situation clear in words,

$M(x)$ = The intensity of jumps having the same sign as x, and greater then x in absolute value, taken with the opposite sign to that of x.

With this definition, $M(x'') - M(x')$ is always the intensity of jumps whose magnitude lies between x' and x'', provided they have the same sign and $x'' > x'$. For $x = 0$, $M(x)$ has a jump $M(+0) - M(-0) = -\mu$, because $M(-0)$ is precisely the intensity of negative jumps, and $M(+0)$ is (with a minus sign) the intensity of positive jumps. The intensity of jumps between some $x' < 0$ and $x'' > 0$ is given by $M(x'') - M(x') + \mu$ (but, usually, one needs to consider separately jumps of opposite sign).

Remark. One can always assume (and we shall always do so, unless we state to the contrary) that there does not exist a probability concentrated at $x = 0$ (i.e. that one can speak of $F(0)$ without having to distinguish between '+0' and '–0', as we have tacitly done when stating that $M(+0) - M(-0) = -\mu[1 - F(0)] - \mu F(0) = -\mu$). In actual fact, so far as any effect on the process is concerned, a 'jump of magnitude $x = 0$' and 'no jump' are the same thing. Mathematically speaking, an increment of F (and hence of M) at $x = 0$ gives a contribution of zero to the integral in equation 8.9, since the integrand vanishes at that point. Sometimes, however, one may make the convention of including in $N(t)$ occurrences of a phenomenon 'able to give rise to a jump', even if the jump does not take place, or, so to speak, is zero. An example of this arises in the field of motor car insurance: if the process $Y(t)$ of interest is the total compensation per accident occurring before time t, it is quite natural (as well as more convenient and meaningful) to count up all accidents, or, to be technical, all claims arising from accidents, without picking out, and excluding, the occasional case for which the compensation was zero. The same principle applies if the process of interest concerns the number of dead, or injured, or those suffering damage to property, and so on.

Formally, in this case one would merely replace μ (the intensity of the jumps) by $\mu + \mu_0$ (where μ_0 is the intensity of the phenomenon with 'zero jumps'), including in $M(x)$ a jump μ_0 at $x = 0$, and consequently altering $F(x)$ and the characteristic function $\chi(u)$, which would be replaced by a mixture of $\chi(u)$ and 1, with weights μ and μ_0. This would be irrelevant, as it must be, since the product $\mu[\chi(u) - 1]$ remains unchanged, and this is all that really matters.

Recall (from Chapter 6, 6.11.6, equation 6.69) that, in order to obtain expressions in normal form ($\mu_0 = 0$), it is necessary and sufficient that we have $(1/2a)\int_{-a}^{a} \chi(u)\,\mathrm{d}u \to 0$ (as $a \to \infty$). Were the limit to equal $c \neq 0$ (necessarily >0), it would suffice to remove $\chi(u)$ and replace it by $[\chi(u) - c]/(1 - c)$.

In the case of a compound process, consisting of a finite number of simple processes (like that considered in Section 8.1.3, the notation of which we continue to use), we have the following:

the $dM(x)$ are the masses (intensities) μ_k concentrated at the values x_k;
$M(x)$ is the sum of the μ_k corresponding to the x_k lying between x and $+\infty$ if $x > 0$, and to those lying between $-\infty$ and x if $x < 0$ (in this case, with the sign changed);
$F(x)$ is the same sum, but running always from $-\infty$ to x, and normalized (i.e. divided by $\mu = \mu_1 + \mu_2 + \ldots + \mu_n$);
the characteristic function for the jumps is given by $\chi(u) = \sum_k e^{iux_k}\mu_k / \mu$, that of the process by

$$\phi^t(u) = \exp\left\{t\sum_k \mu_k\left(e^{iux_k} - 1\right)\right\}, \tag{8.12}$$

which, as is obvious, can also be obtained as the product of the characteristic functions of the superimposed simple processes; that is of the $\exp\left\{t\mu_k\left(e^{iux_k} - 1\right)\right\}$.

Equation 8.10 has the same interpretation in the case of an arbitrary compound Poisson process: it reveals it to be a mixture of simple processes, but no longer necessarily a finite mixture.

Finally, we observe that the prevision $\mathbf{P}[Y(t)]$, and the variance $\boldsymbol{\sigma}^2[Y(t)]$, both exist and are determined by the distribution (both in the strict sense, and in terms of $\hat{\mathbf{P}}$; see Chapter 6, 6.5.7), so long as the same holds true for the jumps; that is if $\mathbf{P}(X)$ and $\boldsymbol{\sigma}^2(X)$, respectively, exist (in the same sense). In this case, we have

$$\mathbf{P}\left[Y(t)\right] = \mu t\mathbf{P}(X), \qquad \boldsymbol{\sigma}^2\left[Y(t)\right] = \mu t\boldsymbol{\sigma}^2(X)$$

(where $\mathbf{P}$ can be replaced by $\hat{\mathbf{P}}$, provided we do so on both sides).

If the prevision makes sense, it also makes sense to consider the process minus the prevision; in other words, modified by subtracting the certain linear function $\mu t\mathbf{P}(X)$, so that we obtain a process with zero prevision (i.e. a finite process). In other words, one considers $Y(t) - \mu t\mathbf{P}(X)$, which is the amount by which $Y(t)$ exceeds the prevision, as in the simple case of equation 8.6. The characteristic function is also similar to that of the latter case, and has the form

$$\phi^t(u) = \exp\left\{\mu t\int(e^{iux} - 1 - iux)\,dF(x)\right\} = \exp\left\{t\int\left(e^{iux} - 1 - iux\right)dM(x)\right\}. \tag{8.13}$$

8.2.8. We are now in a position to characterize the most general form of homogeneous process with independent increments; that is to say, the most general infinitely divisible distribution (see Chapter 6, 6.11.3 and 6.12). In fact, in either formulation it is a question of characterizing the characteristic functions $\phi(u)$ for which $[\phi(u)]^t$ turns out to be a characteristic function for any arbitrary $t > 0$.[6]

We have already encountered an enormous class of infinitely divisible characteristic functions; those of the form $\phi(u) = \exp\{a[\chi(u) - 1]\}$, where $\chi(u)$ is a characteristic function

---

6 For any $t < 0$, this is impossible (except in the degenerate case $\phi(u) = e^{iua}$, in which $|\phi(u)| = 1$; in other cases, for some u we have $|\phi(u)| < 1$, and then, for t 0, we would have $|\phi(u)|^t > 1$). Moreover, it is sufficient to verify the condition for the sequence $t = 1/n$ (or any other sequence tending to zero), rather than for all t. In fact, it holds for all multiples, and hence for an everywhere dense set of values; by the continuity property of Chapter 6, 6.10.3, it therefore holds for all $t > 0$.

(the $\phi^t(u)$ of the compound Poisson processes). A function that is a limit of characteristic functions of this kind (in the sense of uniform convergence in any bounded interval) is again an infinitely divisible characteristic function. To see this, note that the limit of characteristic functions is again a characteristic function, and if some sequence of $\phi_n(u)$, such that $\phi_n \to \phi$, are infinitely divisible, then ϕ_n^t is a characteristic function, $\phi_n^t \to \phi^t$, and hence ϕ^t is a characteristic function (for every $t > 0$); in other words, ϕ is infinitely divisible. Conversely, it can be shown that an infinitely divisible characteristic function is necessarily either of the 'compound Poisson' form, or is a limit case; that is *the set of infinitely divisible characteristic functions coincides with the closure of the set of characteristic functions of the form* $\phi(u) = \exp\{a[\chi(u) - 1]\}$, *where* $\chi(u)$ *is a characteristic function*.

In order to prove this, it is sufficient to observe that if $[\phi(u)]^t$ is a characteristic function for every t, then $\phi_n(u) = \exp\{n[[\phi(u)]^{1/n} - 1]\}$ is also a characteristic function of the compound Poisson type, and tends to $\phi(u)$ (in fact, we are dealing with the well-known elementary limit $n(x^{1/n} - 1) \to \log x$). The process $\phi^t(u)$ is thus approximated by means of the processes $\phi_n^t(u)$ having, perhaps only apparently (see the *Remark* in Section 8.2.7), intensities $\mu_n = n$, and jump distributions $\chi_n(u) = [\phi(u)]^{1/n}$. More precisely, this 'apparently' holds in the cases we have already considered (the compound Poisson processes) and *actually* holds in the new limit cases which we are trying to characterize. In fact, in the compound Poisson cases, with finite intensity μ, the probability

$$p_t(0) = \mathbf{P}\left[Y(t) = 0\right] = \lim(1/2a)\int_{-a}^{a} \phi^t(u)\,\mathrm{d}u \qquad \left(\text{as } a \to \infty\right)$$

(the mass concentrated at 0 in the distribution having characteristic function $\phi^t(u)$) is $\geqslant e^{-\mu t}$ (which is the probability of no jump occurring before time t).[7] In this case, all the $\chi_n(u) = \phi^{1/n}(u)$ contain a constant term at least equal to $e^{-\mu/n}$ (corresponding to the mass at 0) and the actual intensity, instead of being $\mu_n = n$, is at most $n(1 - e^{-\mu/n}) \sim \mu$ (and it is easily verified that it actually tends to μ, as we might have guessed).

The new cases arise, therefore, when $Y(t) = 0$ has zero probability for every $t > 0$, no matter how small; that is when there is zero probability of $Y(t)$ remaining unchanged during any finite time interval, however small. We must have either a continuous variation, or a variation whose jumps are everywhere dense; that is with infinite intensity. In the approximation we have considered, the μ_n will all actually be equal to n.

These remarks and the treatment to follow are rather informal. We shall subsequently often have occasion to dwell somewhat more closely on certain critical aspects of the problem, but for the more rigorous mathematical developments we shall refer the reader to other works (for example, Feller, Vol. II, Chapter XVII, Section 2).

---

7 We have $p_t(0) > e^{-\mu t}$ if and only if there are jump values having nonzero probabilities (concentrated masses in the distribution whose characteristic function is $\chi(u)$) and some sum of them is 0. For example, in the case of Heads and Tails, values ± 1, we have $1 + (-1) = 0$ (i.e. we can return to 0 after two jumps). If, in this same example, the gains had been fixed at +2 and −3, then it would be possible to return to 0 after 5 jumps ($2 + 2 + 2 - 3 - 3 = 0$) etc. In general, $p_t(0) = \sum_{h=2}^{\infty} \mathbf{P}\left[N(t) = h\right]\mathbf{P}(X_1 + X_2 + \ldots + X_h = 0)$, where $\mathbf{P}[N(t) = h) = e^{-\mu t}(\mu t)^h/h!$, and $\mathbf{P}(X_1 + X_2 + \ldots + X_h = 0)$ is the mass concentrated at 0 in the distribution whose characteristic function is $[\chi(u)]^h$.

8.2.9. As a first step in getting to grips with the general case, let us begin by extending to this case the considerations concerning the distribution of the intensity of the jumps, $M(x)$, as defined in Section 8.2.7 for the compound Poisson process. A definition which (making the previous considerations more precise) would be equivalent and which is also directly applicable to the general case is the following: $M(x)$ (taken with a plus or minus sign opposite to that of x) is the prevision of the number of increments having the same sign as x, and greater than x in absolute value, that occur in a unit time interval (subdivided into a large number of very small time intervals). More simply, and more concretely, we shall restrict ourselves to considering the subdivision into n small subintervals, each of length $1/n$, subsequently passing to the limit as $n \to \infty$.

The increment of $Y(t)$ in any one of these subintervals,

$$Y\left(t+1/n\right)-Y\left(t\right),$$

has distribution function $F^{1/n}(y)$ (see Section 8.1.3). The probability that it is greater than some positive x is $1 - F^{1/n}(x)$ and the prevision of the number of increments greater than x is $n[1 - F^{1/n}(x)]$, or, if one prefers, $[1 - F^t(x)]/t$. Similarly, for increments 'exceeding' some negative x (i.e. negative, and greater in absolute value), the probability and prevision are given by $F^t(x)$ and $F^t(x)/t$ ($t = 1/n$). We define $M(x)$ as the limit (as $t = 1/n \to 0$) of $-[1 - F^t(x)]/t$ for positive x, and of $F^t(x)/t$ for negative x. Assuming (as is, in fact, the case) that these limits exist, we can say, to a first approximation, that, for $t \sim 0$, we have $F^t(x) = 1 + tM(x)$ (for $x > 0$) and $F^t(x) = tM(x)$ (for $x < 0$). In other words (in a unified form),

$$F^t\left(x\right)=F^0\left(x\right)+tM\left(x\right),$$

where $F^0(x)$ (the limit case for $t = 0$) represents the distribution concentrated at the origin ($F^0(x) = 0$ for $x < 0$, and $= 1$ for $x > 0$).

This agrees intuitively with the idea that $M(x)$ is the intensity of the jumps 'exceeding' x, and, in particular, in the compound Poisson case, with $M(x) = \mu[F(x) - F^0(x)]$. In the general case, the meaning is the same, except that $M(-0)$ and $M(+0)$ can become infinitely large (either $M(-0) = +\infty$, or $M(+0) = -\infty$, or both), as shown in Figure 8.2.

The passage to the limit, which enables one to obtain the generalized Poisson processes, thus reduces to the construction of the $\phi^t(u)$ on the basis of formulae 8.10 and 8.13 of Section 8.2.7, allowing the function $M(x)$ to become infinite as $x \to \pm 0$, along with appropriate restrictions to ensure that the function converges, and that the process it represents makes sense (but we limit ourselves here to simply indicating how this can be done, and that it is, in fact, possible).

8.2.10. A new form, intermediate between the two previous forms in so far as it provides compensation only for the small jumps, proves more suitable as a basis for a unified account. This is defined by equation 8.14 below, and has to be constructed (within largely arbitrary limits) in such a way that it turns out to be equivalent to equation 8.13 in the neighbourhood of $x = \pm 0$, and to equation 8.10 in the neighbourhood of $x = \pm\infty$. We consider

$$\phi^t\left(u\right)=\exp\left\{t\int\left[\mathrm{e}^{\mathrm{i}ux}-1-\mathrm{i}ux.\tau\left(x\right)\right]\mathrm{d}M\left(x\right)\right\}, \tag{8.14}$$

where $\tau(x)$ is an arbitrary bounded function, tending to 1 as $x \to 0$, and to 0 as $x \to \pm\infty$.

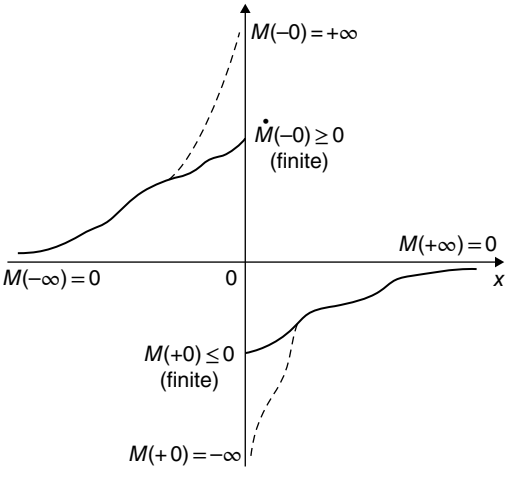


Figure 8.2 Distribution of the intensity of the jumps.

Possible choices are :

$$\tau(x) = (|x| < 1) \quad (\text{i.e. } 1 \text{ in}[-1,1], \text{ and } 0 \text{ else where; P.Lévy}),$$

or

$$\tau(x) = 1/(1 + x^2) \quad (\text{Khintchin}),$$

or

$$\tau(x) = \sin x/x \quad (\text{Feller}).^{8}$$

A necessary and sufficient condition for the expression 8.14 (with any $\tau(x)$ whatsoever) to make sense as the characteristic function of a random process – and, in this way, to provide all the infinitely divisible distributions, apart from the normal, which derives from it as a limit case – is that the contribution to the variance from the 'small jumps' be finite. In other words, we must have $\int x^2 \mathrm{d}M(x) < \infty$ (the integral being taken, for example, over [–1, 1]; the actual interval does not matter, provided it is finite, and contains the origin).[9] We note that, in the more regular case in which the intensity admits a density $M'(x)$, and in which it makes sense to speak of an 'order of infinity' as it tends to 0 (from the left or right), the necessary and sufficient condition is that this order of infinity (from both sides) be <2: if

$$M'(x) \sim 1/x^{\alpha} \ (\alpha < 2),$$

things go through; if $M'(x) \sim 1/x^2$, they no longer do.

---

8 Some authors prefer, in place of $\mathrm{d}M(x)$ as the differential element in the integral, to adopt variants like $\mathrm{d}K(x) = [x/(1 + x^2)]\mathrm{d}M(x)$ (Khintchin), or $\mathrm{d}H(x) = x\,\mathrm{d}M(x)$ (Feller). These have distinct formal advantages, but do not seem to me to compensate for the loss of direct meaning (see Feller, Vol. II, p. 536 *et passim*, and P. Lévy (1965), p. 141).

9 See P. Lévy and Feller, *loc. cit.*

Having said this, it is now easy to make precise the circumstances under which, and the reasons why, the expression in equation 8.14 can be replaced by one or other of the two simpler forms given previously. These can be considered as special cases of equation 8.14, in which $\tau(x)$ (instead of satisfying the imposed conditions) is set equal to 0 for equation 8.10 (the term iux is omitted), and equal to 1 for equation 8.13 (the term iux is always present).

The term iux is innocuous (it merely produces an addition to $Y(t)$ of a certain linear function ct) so long as it is applied to jumps that are neither too large nor too small (e.g. for $\varepsilon < |x| < 1/\varepsilon$, with $\varepsilon > 0$ arbitrarily small). When applied in a neighbourhood of $x = 0$, it is either innocuous or *useful;* when applied in a neighbourhood of ∞ (e.g. for $1/|x| < \varepsilon$), it is either innocuous or *harmful*. It can be useful, and indeed *necessary*, when the *small jumps*, if not 'compensated', do not lead to convergence; this happens if $\int |x| \, dM(x)$ diverges in $[-1, 1]$ (or, equivalently, in any neighbourhood of 0). It can be harmful, because the 'compensation' of the '*big jumps*' destroys convergence if they give too large a contribution, and this happens precisely when the previous integral diverges over $|x| \geqslant 1$. Observe, also, that this may very well happen even in a compound Poisson process (μ finite); it is enough that the distribution of the jumps should fail to have a 'mean value' (as, for example, with the Cauchy distribution). This does not affect the process, but any attempt to pass to the limit for the 'compensated' jumps would destroy the convergence rather than assist it.

In conclusion: the condition $\int x^2 dM(x) < \infty$ over $[-1, 1]$ is necessary and sufficient, and expression 8.14 always holds if the condition is satisfied. Both of the simple forms equations 8.10 and 8.13 can be applied if $\int |x| dM(x) < \infty$ over $[-\infty, +\infty]$ (and we observe that this condition implies the general condition and is, therefore, itself sufficient). If, instead, the integral diverges, it is necessary to distinguish whether this is due to contributions in the neighbourhood of the origin, or in the neighbourhood of $\pm\infty$; in the first case equation 8.10 is ruled out, and in the second equation 8.13 is ruled out (and both are if there is trouble both at the origin and at infinity).

8.3 The Wiener–Lévy Process

8.3.1. We now turn to an examination of the continuous component of a homogeneous random process with independent increments, which we described briefly in Section 8.1.3: this is the Wiener–Lévy process. The points already made, together with some further observations, will suffice here to provide a preliminary understanding of the process and will be all that is required for the discussion to follow.

Let us first make clear what is meant by calling the process 'continuous'. It amounts to saying that, for any preassigned $\varepsilon > 0$, if we consider the increments of $Y(t)$ in $0 \leqslant t \leqslant 1$, divided up into N equal intervals, the probability that even one of the increments is, in absolute value, greater than ε, that is the possibility that $|Y(t + 1/N) - Y(t)| > \varepsilon$ for some $t = h/N < 1$, tends to 0 as N increases. In short, we are dealing with 'what escapes the sieve laid down for the selection of the jumps'. There is no harm in thinking (on a superficial level) of this as being equivalent to the continuity of the function $Y(t)$. From a conceptual point of view, however, this would be a distortion of the situation, as can be seen from the few critical comments we have already made, and from the others we shall be presenting later, albeit rather concisely, in a more systematic form (see, for example, Section 8.9.1; especially the final paragraph).

We shall usually deal with the *standardized* Wiener–Lévy process, having zero prevision, $m(t) = 0$, unit variance per unit time period, $\sigma^2(t) = t$, and initial condition $Y(0) = 0$.

In this case, the density $f^t(y)$ and the characteristic function $\phi^t(u)$ (both at time t) are given by

$$f^t(y) = K_t^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}y^2/t} \quad \left(K = 1/\sqrt{(2\pi)}\right), \tag{8.15}$$

$$\phi^t(u) = \mathrm{e}^{-\frac{1}{2}tu^2}. \tag{8.16}$$

The general case ($Y(0) = y_0$, $m(t) = mt$, $\sigma^2(t) = t\sigma^2$) can be reduced to the standardized form by noting that it can be written as $a + mt + \sigma Y(t)$, where $Y(t)$ is in standardized form. If we wish to consider it explicitly in the general form, we have

$$f^t(y) Kt^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left[(y-yo-mt)/\sigma\right]_{2/t}} \quad \left(K = 1/\sqrt{(2\pi)}\sigma\right),$$

$$\phi^t(u) = \mathrm{e}^{\mathrm{i}uyo}.\mathrm{e}^{t\left(\mathrm{i}um - \frac{1}{2}u^2\sigma^2\right)}$$

(where, for greater clarity, the terms depending on the initial value, y_0, and those depending on the process, i.e. on t, are written separately).

8.3.2. The same approach holds good when we wish to examine the random process and its behaviour, rather than just isolated values assumed by it. In fact, the joint distribution of the values of $Y(t)$ at an arbitrary number of instants $t_1, t_2 \ldots, t_n$, is also a normal distribution, with density![10] given by

$$f\left(y_1, y_2, \ldots, y_n\right) = K \quad \mathrm{e}^{-\frac{1}{2}Q\left(y_1, y_2, \ldots, y_n\right)},$$

where Q is a positive definite quadratic form determined by the covariances $\mathbf{P}[Y(t_i)Y(t_j)] = \sigma_i\sigma_j r_{ij}$ (if $i = j$, $r_{ii} = 1$, and covariance = variance = σ_i^2). We have already seen (in Section 8.1.6) that (if $t_i \leqslant t_j$) the covariance is t_i^2, and therefore $r_{ij} = \sqrt{(t_i/t_j)}$; this gives all the information required for any application.

It is simpler and more practical, however, to observe that all the $Y(t_i)$ can be expressed in terms of increments, $\Delta_i = Y(t_i) - Y(t_{i-1})$, which follow consecutively and are independent: $Y(t_i) = \Delta_1 + \Delta_2 + \ldots + \Delta_i$. But Δ_i has a centred normal distribution, with variance $(t_i - t_{i-1})$, and Q, as a function of the variables $(y_i - y_{i-1})$, is a sum of squares:

$$Q\left(y_1, y_2, \ldots, y_n\right) = \sum_i \left[\left(y_i - y_{i-1}\right)^2 / \left(t_i - t_{i-1}\right)\right]. \tag{8.17}$$

We shall now make use of this to draw certain conclusions, which we shall require in what follows.

---

10 We assume that $m = 0$, $\sigma = 1$ (the standardized case); only trivial modifications are required for the general case.

What we have been considering so far, to be absolutely precise, is the Wiener–Lévy process on the half-line $t \geqslant 0$, given that $Y(0) = 0$; the case of $t \geqslant t_0$, given that $Y(t_0) = y_0$, is identical (and the same is true for $t \leqslant 0$, $t \leqslant t_0$, respectively).[11]

In order to consider the case in which several values are given (at $t = t_1, t_2, \ldots$), it is sufficient to consider the problem inside one of the (finite) intervals; on the unbounded intervals things proceed as above. Let us, therefore, consider the process over the interval (t_1, t_2), given the values $Y(t_1) = y_1$ and $Y(t_2) = y_2$ at the end-points. In order to characterize it completely, it will suffice, in this case also, to determine the prevision (no longer necessarily zero!) and the variance of $Y(t)$ for each $t_1 \leqslant t \leqslant t_2$, and the covariance (or correlation coefficient) between $Y(t')$ and $Y(t'')$ for each pair of instants $t_1 \leqslant t' \leqslant t'' \leqslant t_2$.

We now decompose $Y(t)$ into the sum of a certain linear component (a straight line through the two given points) and the deviation from it:

$$Y(t) = y_1 + \left[(t - t_1)/(t_2 - t_1)\right](y_2 - y_1) + Y_0(t),$$

where $Y_0(t)$ corresponds to the same problem with $y_1 = y_2 = 0$. We now consider the two components as if the end-points were not yet fixed, so that the increments $\Delta_1 = Y - Y_1$ and $\Delta_2 = Y_2 - Y$ are independent, with standard deviations $\sigma_1 = \sqrt{(t - t_1)}$ and $\sigma_2 = \sqrt{(t_2 - t)}$. The linear component is then the random quantity

$$\left[1/(t_2 - t_1)\right]\left[(t_2 - t)Y_1 + (t - t_1)Y_2\right],$$

and, by subtraction, $Y_0(t)$ is given by

$$Y_0(t) = \left[1/(t_2 - t_1)\right]\left[(t_2 - t)\Delta_1 - (t - t_1)\Delta_2\right].$$

It is easily seen that $Y_0(t)$ has zero prevision (as was obvious) and standard deviation given by

$$\sigma\left(Y_0(t)\right) = \sqrt{\left[(t - t_1)(t_2 - t)/(t_2 - t_1)\right]}; \tag{8.18}$$

moreover, it is uncorrelated with the linear component

$$Y_1 + \left[(t - t_1)/(t_2 - t_1)\right](\Delta_1 + \Delta_2)$$

(and hence, by normality, they are independent). In fact, we have

$$\begin{aligned} \text{covariance} &= K\left[(t - t_1)(t_2 - t)\mathbf{P}\left(\Delta_1^2\right) - (t - t_1)^2\,\mathbf{P}\left(\Delta_2^2\right)\right] \\ &= K\left[(t_2 - t)\sigma_1^2 - (t - t_1)\sigma_2^2\right] = 0. \end{aligned}$$

It can be seen (as a check, and in order to realize the difference between this and $\sqrt{(t - t_1)}$, which would apply in the absence of the condition on the second end-point) that the

11 Provided that the process is assumed to make sense, even in the past, and that, in fact, no knowledge of the past leads us to adopt different previsions (these assumptions are not, in general, very realistic).

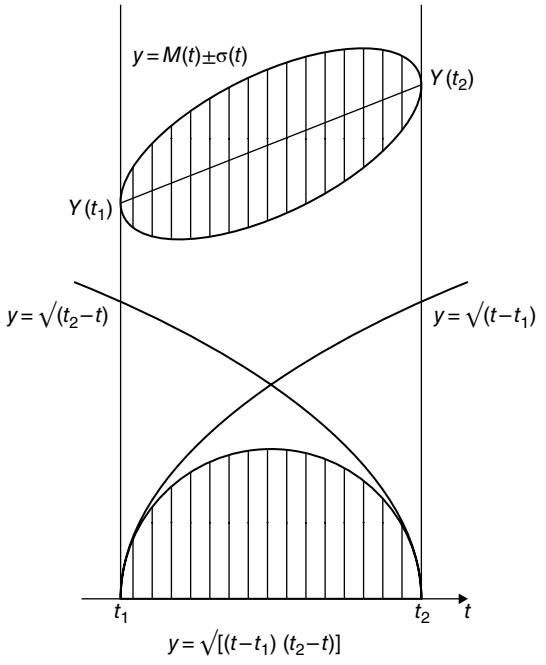


Figure 8.3 Interpolation between two known points in the Wiener–Lévy process (straight line and semi-ellipses: graphs of *prevision* and *prevision* ± *standard deviation*). The lower diagram represents the behaviour of the standard deviation given the point of origin, the final point, or both.

standard deviation of the linear component (considering the value of the first end-point, Y_1, as fixed) is given by

$$\left[\left(t-t_1\right)/\left(t_2-t_1\right)\right]\sqrt{\left[\mathbf{P}\left(\Delta_1+\Delta_2\right)^2\right]}=\sqrt{\left(t-t_1\right)}.\sqrt{\left[\left(t-t_1\right)/\left(t_2-t_1\right)\right]}.$$

Summing the squares of the standard deviations of the two summands, we obtain the square of $\sigma_1 = \sqrt{(t - t_1)}$ (as, indeed, we should). It is useful to note that, as is shown in Figure 8.3, the standard deviation (equation 8.18) of $Y(t)$ given the values at t_1 and t_2, is represented by the (semi) circle resting on the segment (t_1, t_2) (provided the appropriate scale is used; i.e. taking the segment $t_2 - t_1$ on the t-axis equal to the unit of measure on the y-axis: on the other hand, this is really irrelevant, except as an aid to graphical representation and description). If we consider the parabolae that represent, in a similar fashion, $\sigma(t)$ given only Y_1 (i.e. $y = \sqrt{(t - t_1)}$), or given only Y_2 (i.e. $y = \sqrt{(t_2 - t)}$), we see that the product of these two functions is represented by the circle, which, therefore, touches the parabolae (at the end-points), because when one of the two factors vanishes, the other has value 1.

We can determine, in a similar manner, the covariance between

$$Y\left(t'\right)\text{and}\,Y\left(t''\right),\quad t_1\le t'\le t''\le t_2.$$

We denote the successive, independent increments by

$$\Delta_1=Y\left(t'\right)-Y\left(t_1\right),\quad \Delta_2=Y\left(t''\right)-Y\left(t'\right),\quad \Delta_3=Y\left(t_2\right)-Y\left(t''\right);$$

$Y_0(t)$ will be the same as before, but writing $y_2 - = y_1 = \Delta_1 + \Delta_2 + \Delta_3$, $T = t_2 - t_1$ and assuming, simply for notational convenience, that $t_1 = 0$ and $y_1 = 0$, we have

$$Y_0(t') = \Delta_1 \qquad -t'(\Delta_1 + \Delta_2 + \Delta_3) = (T - t')\Delta_1 \qquad -t'\Delta_2 - t'\Delta_3,$$
$$Y_0(t'') = \Delta_1 + \Delta_2 - t''(\Delta_1 + \Delta_2 + \Delta_3) = (T - t'')\Delta_1 + (T - t'')\Delta_2 - t''\Delta_3,$$

and hence

$$Y_0(t')Y_0(t'') = (T - t')(T - t'')\Delta_1^2 - t'(T - t'')\Delta_2^2 + t't''\Delta_3^2$$
+ cross-product terms (in $\Delta_i\Delta_j$, $i \neq j$, with zero prevision).

Taking prevision, and bearing in mind that the previsions of the Δ_i^2 are, respectively, $\sigma_1^2 = t'$, $\sigma_2^2 = t'' - t'$, $\sigma_3^2 = T - t''$ we have

$$\begin{aligned}\operatorname{Convar}(t',\ t'') &= (T - t')(T - t'')t' - t'(T - t'')(t'' - t') + t't''(T - t'') \\ &= t'(T - t'')\left[(T - t') - (t'' - t') + t''\right] = t'(T - t'')T.\end{aligned}$$

Dividing by $\sigma' = \sqrt{\left[t'(T - t')/T\right]}$ and $\sigma'' = \sqrt{\left[t''(T - t'')/T\right]}$, we obtain, finally, the correlation coefficient

$$r(t', t'') = \sqrt{\left[\frac{t'(T - t'')}{(T - t')t''}\right]} = \sqrt{\left[\frac{(t' - t_1)(t_2 - t'')}{(t_2 - t')(t'' - t_1)}\right]}, \tag{8.19}$$

and, in this way, we return to our original notation.

8.3.3. It will certainly be no surprise to find that if we put $t_2 = \infty$ we reduce to the results obtained in the case of the single condition at t_1. This is, however, simply a special case of a remarkable fact first brought to light by P. Lévy, and which is often useful for inverting this conclusion by reducing general cases (with two fixed values) to the special limit case (with only one fixed value). The fact referred to is the *projective invariance* of problems concerning processes of this kind,[12] and derives from the expression under the square root in $r(t', t'')$ being the cross-ratio of the four instants involved. It therefore remains invariant under any homographic substitution for the time t, provided the substitution does not make the *finite* interval $[t_1, t_2)$ correspond to the complement of a finite interval (but instead, to either a finite interval, or to a half-line; in other words, the inequalities $t_1 \leqslant t' \leqslant t'' \leqslant t_2$ must either be all preserved, or all inverted). Consequently, the stochastic nature of the function $Y(t)$ remains invariant if we ignore multiplication by a certain arbitrary function; invariance holds for any random function of the form $Z(t) = g(t)Y(t)$, or, in particular, for the standardized process (always having $\sigma = 1$) that can be obtained by taking

$$g(t) = 1/\sigma\left[Y(t)\right] = 1/\sqrt{t}. \tag{8.20}$$

---

12 The meaning of this is most easily understood by introducing the projective coordinate

$$\tau = \tau(t) = (t - t')(t'' - t_2)/(t'' - t)(t_2 - t');$$

i.e. (as is obvious) taking t', t_2 and t'' to 0, 1 and ∞ : $\tau' = \tau(t') = 0$, $\tau_2 = \tau(t_2) = 1$, $\tau'' = \tau(t'') = \infty$.
It follows that $r = \sqrt{\tau_1}$, where $\tau_1 = \tau(t_1)$ is the abscissa of the point t_1 after the projective transformation has been performed.

This device will prove useful for, among other things, reducing the study of the asymptotic behaviour of $Y(t)$ in the neighbourhood of the origin (for $0 < t < \varepsilon$, $\varepsilon \to 0$) to that at infinity (for $t > T$, $T \to \infty$); see Section 8.9.7.

The basic properties of the Wiener–Lévy process will be derived later, in contexts where they correspond to actual problems of interest.

8.4 Stable Distributions and Other Important Examples

8.4.1. We have already encountered (in Chapter 6, 6.11.3) two families of stable distributions: the normal (which, in Chapter 7, 7.6.7, we saw to be the only stable distribution having finite variance) and the Cauchy (which has infinite variance).

We are now in a position to determine all stable distributions. It is clear – and we shall see this shortly – that they must be infinitely divisible. Our study can, therefore, be restricted to a consideration of distributions having this latter property and, since we know what the explicit form of their characteristic functions must be, this will not be a difficult task.

Knowledge of these new stable distributions will also prove useful for clarifying the various necessary conditions and other circumstances occurring in the study of the asymptotic behaviour of random processes.

We begin by observing that the convolution of two compound or generalized Poisson distributions, represented by the distributions of the intensities of the jumps, $M_1(x)$ and $M_2(x)$, is obtained by summing them: $M(x) = M_1(x) + M_2(x)$. In fact, $M(x)$ determines the logarithm of the characteristic function in a linear fashion, and the sum in this case corresponds to the product of the characteristic functions; that is to the convolution.

This makes clear, conversely, what the condition for an infinitely divisible distribution to be a factor in the decomposition of some other distribution (also infinitely divisible) must be. The distribution defined by $M_1(x)$ 'divides' the distribution defined by $M(x)$ if and only if the difference

$$M_2(x) = M(x) - M_1(x)$$

is also a distribution function of intensities. This implies that it must never decrease, so that every interval receives positive mass (or, at worst, zero mass), and this implies, simply and intuitively, that in any interval (of the positive or negative semi-axis) $M_1(x)$ must have an increment not exceeding that of $M(x)$. In particular, the concentrated masses and the density (if they exist) must, at every individual point, not exceed those of $M(x)$. If one wishes to include in the statement the case in which a normal component exists (and then we have the most general infinitely divisible distribution) it is sufficient to state that here, too, this component must be the smaller (as a measure, one can take the variance).[13]

In order to prove that stability implies infinite divisibility, it is sufficient to observe that, in the case of stability, the sum of n independent random quantities that are

13 Note that what we have said concerns divisibility within the class of infinitely divisible distributions. However, there may exist indivisible factors of infinitely divisible distributions (and, of course, conversely), as we mentioned already in Chapter 6, 6.12.

identically distributed has again the same distribution (up to a change of scale). It is itself, therefore, a convolution product of an arbitrary number, n, of identical factors, and is therefore infinitely divisible. If, for each factor, the distribution of the intensities of the jumps is $M(x)$, then for the convolution of n factors it is $nM(x)$.

For stability it is necessary and sufficient that the distribution defined by $nM(x)$ belongs to the same family as that defined by $M(x)$. This requires that it differs only by a (positive) scale factor $\lambda(n)$: that is $nM(x) = M(\lambda(n)x)$. It follows immediately that

$$knM(x) = kM(\lambda(n)x) = M(\lambda(k)\lambda(n)x) = M(\lambda(kn)x);$$

in other words, $\lambda(k)\lambda(n) = \lambda(kn)$ for k and n integer. The same relation also holds for rationals if we set $\lambda(1/n) = 1/\lambda(n)$ and hence $\lambda(m/n) = \lambda(m)/\lambda(n)$. By continuity, we then have $\lambda(\nu)$ for all positive reals ν. The functional equation $\lambda(\nu_1)\lambda(\nu_2) = \lambda(\nu_1\,\nu_2)$ characterizes powers, so we have an explicit expression for λ:

$$\lambda(\nu) = \nu^{-1/\alpha}; \qquad \text{in particular}, \lambda(n) = n^{-1/\alpha}. \tag{8.21}$$

We have written the exponent in the form $-1/\alpha$, because it is the reciprocal, $-\alpha$, which appears as the exponent in the expression for $M(x)$, and which will be of more direct use in what follows. It is for this reason that α is known as the 'characteristic exponent' of the distribution for which we have

$$nM(x) = M\left(n^{-1/\alpha}x\right) \quad \left(\text{and, in general, } \nu M(x) = M\left(\nu^{-1/\alpha}x\right), 0 < \nu < +\infty\right). \tag{8.22}$$

In fact, we can immediately obtain an explicit expression for $M(x)$. When $x = 1$, the above expression reduces to $\nu M(1) = M(\nu^{-1/\alpha})$, and when $x = \nu^{-1/\alpha}$, we have $M(x) = -Kx^{-\alpha}$, where $-K = M(1)$, a constant: this holds for all positive x (running from 0 to $+\infty$ as ν varies in the opposite direction from $+\infty$ to 0). Setting $x = -1$ (instead of $+1$), we can obtain the same result for negative x, except that we must now write $|x|^{-\alpha}$ in place of $x^{-\alpha}$. Allowing for the fact that the constant K could assume different values on the positive and negative semi-axes, we have, finally,

$$M(x) = -K^{+}|x|^{-\alpha}(x > 0) + K^{-}|x|^{-\alpha}(x < 0) = K^{\pm}|x|^{-\alpha}, \tag{8.23}$$

where K^{+} and K^{-} are positive, and are therefore written preceded by the appropriate sign (this ensures that $M(x)$ is increasing, in line with what we said in Section 8.2.9). It is obviously unnecessary to write $|x|$ when x is positive, but it serves to stress the identical nature of the expressions over the two semi-axes: $K^{\pm}$ is merely a compact way of writing either $-K^{+}$ or $+K^{-}$ for $x \gtrless 0$ (it could be written in the form $K^{\pm} = K^{-}(x < 0) - K^{+}(x > 0)$).

It remains to examine which values are admissible for the characteristic exponent, We see immediately that these are the α for which $0 < \alpha \leqslant 2$, and it turns out that there are good reasons for considering these as four separate subcases;

$$0 < \alpha < 1, \qquad \alpha = 1, \qquad 1 < \alpha < 2, \qquad \alpha = 2.$$

The value $\alpha = 1$ is a somewhat special case, and corresponds to the Cauchy distribution (we have already met this in Chapter 6, 6.11.3; the correspondence is established by examining the characteristic function given in equation 6.59 of the section mentioned).

8.4.2. For $\alpha = 2$, we cannot proceed in the above manner – the expression for the characteristic function diverges – but we can consider it as a limit case (or we could include it by using the kind of procedures mentioned in the footnote 8). The limit case turns out to be the by now familiar normal distribution.

In fact, this corresponds to the characteristic exponent

$$\alpha = 2 \qquad \left(\text{or}, 1/\alpha = \tfrac{1}{2}\right)$$

because the scale factor (in this case, the standard deviation) for the sum Y_n of n identically distributed summands is multiplied by $\sqrt{n}$, that is $n^{½}$ (and hence, for the mean Y_n/n we multiply by $n^{-½}$).

More generally, even in the case of distributions from the same family but with different scale factors, the well-known formula for the standard deviation (for sums of independent quantities) holds for the normal distribution,

$$\sigma = \left(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2\right)^{\frac{1}{2}}, \tag{8.24}$$

and also for all stable distributions if adapted to the appropriate characteristic exponent α. Explicitly, if $X_1, X_2, \ldots, X_n$ have distributions all belonging to the same family (stable, with characteristic exponent α), and, $a_1, a_2, \ldots, a_n$ are the respective scale factors, then the random quantity defined by the sum $aX = a_1X_1 + a_2X_2 + \ldots + a_nX_n$ again has a distribution belonging to this family, with scale factor given by

$$a = \left(a_1^\alpha + a_2^\alpha + \ldots + a_n^\alpha\right)^{1/\alpha} \qquad \left(i.e. \quad a^\alpha = \sum a_i^\alpha\right). \tag{8.25}$$

In particular, if all the a_i are equal to 1,

$$a = n^{1/\alpha}. \tag{8.26}$$

This is an immediate consequence of the expression for $M(x)$ given in equation 8.23; setting $\nu = \nu_1 + \nu_2$ in $\nu M(x) = M(\nu^{-1/\alpha}x)$, we obtain

$$M\left(\nu^{-1/\alpha}x\right) = M\left(\nu_1^{-1/\alpha}x\right) + M\left(\nu_2{}^{-1/\alpha}x\right).$$

We began with the case $\alpha = 2$, not only because of its importance and familiarity but also because it enables us to establish immediately that values $\alpha > 2$ are not admissible. This is not only because, *a fortiori*, the integral would diverge; there is another, elementary, or, at least, familiar, reason (which we shall just mention briefly). If the variance is finite, the formula for standard deviations holds when $\alpha = 2$; if it is infinite, we must have $\alpha \leqslant 2$, because $\alpha = 2$ holds for any bounded portion of the distribution (considering, for example, the *truncated* X_h; $-K \vee X_h \wedge K$).

In connection with the idea of *compensation* (e.g. for errors of measurement), there is, by virtue of the (magic?) properties of the arithmetic mean, a point of some conceptual and practical importance which is worth making. (Of course, we are dealing with a mathematical property which we would have had to mention anyway, in order to deal with an important aspect of the behaviour of the means Y_n/n, of n summands, each of which follows a stable distribution with some exponent α.)

The form given in equation 8.26 asserts that, compared with the individual summands, Y_n has scale factor $n^{1/\alpha}$; it follows, therefore, that, for the arithmetic mean Y_n/n, the scale factor is $n^{(1/\alpha)-1}$.

Taking, for example, the factor is

$\alpha =$	2	3/2	4/3	1	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$=$	$n^{-\frac{1}{2}}$	$n^{-\frac{1}{3}}$	$n^{-\frac{1}{4}}$	$\mathbf{1}$	$n^{\frac{1}{4}}$	n	n^3

giving, for example, for

$n=2$	$0\cdot707$	$0\cdot793$	$0\cdot841$	1	$1\cdot189$	2	8
$n=10$	$0\cdot316$	$0\cdot464$	$0\cdot681$	1	$1\cdot779$	10	1000

The usual amount of 'compensation' (i.e. an increase in precision in the ratio 1: $\sqrt{n}$ for the mean of n values) is only attained for $\alpha = 2$; the increase diminishes as α moves from 2 to 1, and for $\alpha = 1$ (the Cauchy distribution) the precision is unaltered, implying no advantage (or disadvantage) in using a mean based on several values rather than just using a single value; for $\alpha < 1$, the situation is reversed and worsens very rapidly as we approach zero (a limit value which must be excluded since it would give ∞!). The values given above should be sufficient to provide a concrete numerical feeling for the situation.

One should not conclude, however, that in these latter cases there is no advantage in having more information (this would be a narrow, short-sighted mistake, resulting from the assumption that the only way to utilize repeated observations is by forming their mean). There is always an advantage in having more observations (there is more information!) but, to make the most of it, it is necessary to pose the problem properly, in a form corresponding to the actual circumstances. This kind of problem of mathematical statistics is dealt with by the Bayesian formulation given in Chapters 11 and 12.

8.4.3. For $\alpha < 2$, we have, in fact, a generalized Poisson distribution, with $M(x)$ corresponding to the distribution function of the intensities of the jumps. Moreover, since $M(x) = K^+.|x|^{-\alpha}$, the density exists and is given by

$$M'(x) = \alpha \left| K^{\pm} \right| |x|^{-(\alpha+1)}. \tag{8.27}$$

It is simpler and clearer (apart from an exceptional case which arises for $\alpha = 1$) to consider separately the distribution generated by the positive jumps (the other is symmetric). We then have, taking $K^+ = -1/\alpha$ in order to obtain the simplest form of the density,

$$M(x) = -(1/\alpha)x^{-\alpha}, \quad M'(x) = x^{-(\alpha+1)}, \quad \mathrm{d}M(x) = \mathrm{d}x/x^{\alpha+1}. \tag{8.27′}$$

It is here that we meet the circumstance which forces us to distinguish between the two cases of α less than or greater than 1 (the case $\alpha = 1$ will appear later on). The fact is that in the first case we have convergence without requiring the correction term iux in equation 8.13, whereas in the second case this term is required. The reason for this is (roughly speaking) that $e^{iux} - 1 \sim iux$ is an infinitesimal of the first order in x for $x \sim 0$; if multiplied by $\mathrm{d}M(x) = \mathrm{d}x/x^{\alpha+1}$, it gives $\mathrm{d}x/x^{\alpha}$, and the integral converges or diverges as $x \to 0$, according to whether $\alpha < 1$ or $\alpha > 1$. This is not merely a question of analysis, however; there is a point of substance involved here. For $\alpha < 1$, the generalized Poisson

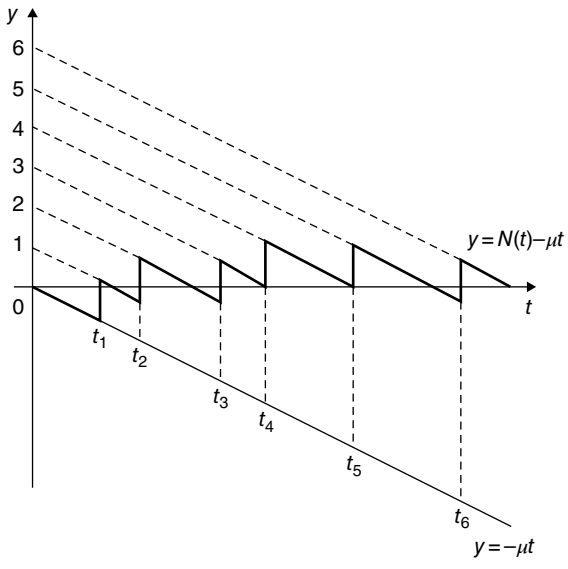


Figure 8.4 The simple, compensated Poisson process (prevision = 0).

random process produced by the positive jumps only, with distribution of intensities $M(x) = Kx^{-\alpha}$ (which is therefore always increasing), makes sense. For $\alpha > 1$, however, only the *compensated sum* of the jumps makes sense. With reference to Figures 8.4 and 8.5,[14] we can give some idea of this behaviour by saying that (as more and more of the numerous very small jumps are added) the sum of the jumps (per unit time) becomes infinite, but the sloping straight line from which we start also becomes infinitely inclined downwards. The process, under these conditions, cannot have monotonic behaviour (in any time interval; no matter how short).

8.4.4. Notwithstanding the diversity of their behaviour, both mathematically and in terms of the actual processes, there are no differences in the form of the characteristic functions. Simple qualitative considerations will suffice to establish that they must have the form $\exp(C|u|^{\alpha})$ (where C is replaced by its complex conjugate, C^*, if u is negative). Detailed calculation (see, for example, **P**. Lévy (1965), p. 163) shows that in the case of positive jumps we have

$$\phi(u) = \exp\left\{-e^{\pm\frac{1}{2}i\pi\alpha}|u|^{\alpha}\right\}, \tag{8.28}$$

14 Figure 8.4 shows a simple Poisson process with its prevision subtracted off (see Section 8.2.6). Figure 8.5 shows what happens in compound processes obtained by superposing, successively, on top of the previous one, simple processes having, in each case, a smaller jump. If one imagines, following on from the three shown, a fourth step, a fifth step, …, and so on, with the slope of the straight line of the prevision increasing indefinitely, one gets some idea of the generalized Poisson processes with only positive jumps.
Note that, in order not to make the diagram too complicated, it has been drawn as if all the additional processes vanish at the instants of the preceding jumps (this is very unlikely, but does not alter the accuracy of the visual impression; we merely wish to warn those who realize that this device has been adopted not to imagine that it reflects some actual property of the processes in question).

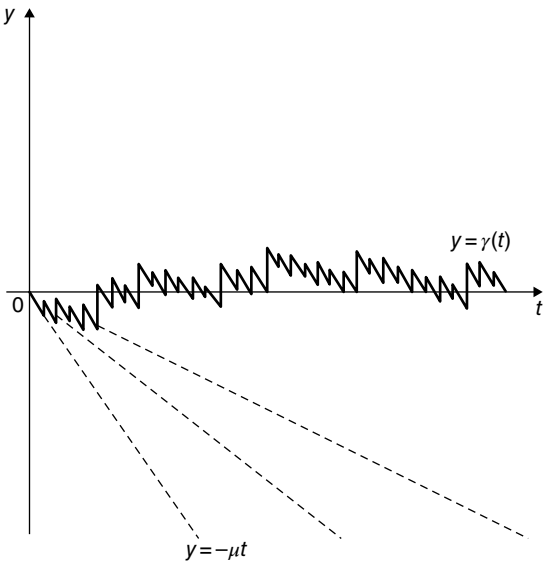


Figure 8.5 The compound Poisson process with successive sums from simple, compensated Poisson processes.

with $\pm$, depending on the sign of u, whereas for negative jumps the $\mp$ signs in the exponent are interchanged. In the general case, it is sufficient to replace α or $-\alpha$ in the exponent by an intermediate value; in particular, in the symmetric case, by 0.

For $\alpha = 1$, the symmetric case gives the Cauchy distribution. The latter can, therefore, be thought of as generated by a generalized Poisson process in which the distribution of the intensities of the jumps is given by

$$M(x) = \pm 1/x,$$

with density $M'(x) = 1/x^2$. In this case, however, we can only ensure convergence by having recourse to the form given in equation 8.14 (this is because the term iux is necessary in the neighbourhood of the origin, but gives trouble at infinity). By doing this, however, we effect a *partial* compensation in the jumps and this reintroduces a certain, arbitrary, additive constant, which prevents the distribution from being stable (following Lévy, we could term it *quasi-stable*; the convolution involves not merely a change of scale, but also a translation). In the symmetric case, we obtain stability by using the same criterion of compensation for the contributions of both negative and positive jumps (or by implicitly compensating, by first integrating between $\pm a$ and then letting $a \to \infty$).

Apart from the two cases $\alpha = 2$ (normal) and $\alpha = 1$ (Cauchy), the densities of stable distributions cannot, in general, be expressed in simple forms (although they exist, and are regular). One exception is the case $\alpha = \frac{1}{2}$, to which corresponds, as an increasing process (positive jumps; $x > 0$),

$$M(x) = -2x^{-\frac{1}{2}}, \quad M'(x) = x^{-\frac{3}{2}}, \quad \mathrm{d}M(x) = \mathrm{d}x / \sqrt{x^3},$$

and the density

$$f(x) = Kx^{-\frac{3}{2}} e^{-1/2x}. \tag{8.29}$$

Finally, we should also mention the case $\alpha = \frac{3}{2}$, which is important on account of an interpretation of it given by Holtsmark in connection with a problem in astronomy (and by virtue of the fact that it precedes knowledge of the problem on the part of mathematicians): the case $\alpha = \frac{4}{3}$ can be given an analogous interpretation (in a four-dimensional world). (See Feller, Vol. II, pp. 170, 215.)

8.4.5. Other important examples of Poisson-type processes are the *gamma* processes (and those derived from them), and those of *Bessel* and *Pascal*.

The gamma distribution (see equations 6.55 and 6.56 in Chapter 6, 6.11.3) has density and characteristic function defined by

$$f^t(x) = Kx^{t-1} e^{-x} \quad (x \geqslant 0), \quad K = \frac{1}{\Gamma(t)}, \tag{8.30}$$

$$\phi^t(u) = 1/(1 - iu)^t. \tag{8.31}$$

As $t \to 0, f^t(x)/t \to e^{-x}/x$, and so the gamma process, in which $Y(t)$ has a gamma distribution with exponent t, derives from jumps whose intensities have the distribution

$$M(x) = \int_x^\infty (e^{-x}/x)\,dx, \quad M'(x) = e^{-x}/x. \tag{8.32}$$

In interpreting this, we note the connection with the Poisson process: $f^t(x)$, for $t = h$ integer, gives the distribution of the waiting-time, T_h, for the hth occurrence of the phenomenon. We can also obtain this by arguing in terms of the $p_h(t)$ (see Section 8.2.5); this makes it completely obvious, because for $h = 1$ we have the exponential distribution (for T_1, and, independently, for any waiting time $T_h - T_{h-1}$) and for an arbitrary integer h we have the convolution corresponding to the sum T_h of the h individual waiting times.

It is interesting to note that one possible interpretation of the gamma process is as the *inverse* of the (simple) Poisson process. To see this, we interchange the notation, writing this process as $t = T(y)$, the inverse of the other, for which we keep the standard notation, $y = Y(t)$. The inverse function $y = T^{-1}(t)$ (which, of course, does not give a process with independent increments), considered only at those points for which y = integer (or taking $Y(t)$ = the integral part of $y = T^{-1}(t)$), gives precisely the simple Poisson process (with $\mu = 1$).

Remark. We obtain a perfect interpretation if we think, for example, of $y = T^{-1}(t)$ as representing the number of turns (or fractions thereof) made by a point moving in a series of jerks around the circumference of a circle. The standard Poisson situation then corresponds to that of someone who is only able to observe the point when it passes certain given marks.

We are in no way suggesting, however, that this mathematical possibility of considering an 'explanation' in terms of some 'hidden mechanism' provides any automatic justification for metaphysical flights of fancy leading on to assertions about its 'existence' (whereas it may, of course, be useful to explore the possibility if there are some concrete reasons for considering it plausible). There are a number of cases (but not, so far as I know, the present one) in which these kinds of metaphysical interpretation (or so they appear to me, anyway) are accepted, or, at any rate, seriously discussed.

More generally, changing the scale and intensity, we have

$$f^t(x) = Kx^{\mu t-1}\mathrm{e}^{-\lambda x}\,(x>0),\ K=\lambda^{\mu t}/\Gamma(\mu t), \tag{8.33}$$

$$\phi^t(\mu) = (1-\mathrm{i}u/\lambda)^{\mu t}. \tag{8.34}$$

Moreover, we can also reflect the distribution onto the negative axis; $f^t(x)$ is unchanged, apart from writing $|x|$ in place of x and changing the final term to $(x < 0)$, and the characteristic function is given by

$$\phi^t(\mu) = (1+\mathrm{i}u/\lambda)^{\mu t}. \tag{8.35}$$

By convolution, we can construct other processes corresponding to sums of gamma processes. The most important case is that obtained by symmetrization (see equations 6.57 and 6.59 in Chapter 6, 6.11.3), which, for $t = 1$, gives the double-exponential distribution. The general case is obtained from products of the form

$$\phi^t(\mu) = (1-\mathrm{i}u/\lambda)^{\mu_1 t}(1-\mathrm{i}u/\lambda_2)^{\mu_2 t}\ldots(1-\mathrm{i}u/\lambda_n)^{\mu_n t}$$

(where the signs can be either – or +; we could, alternatively, say that the λ_h can be positive or negative); the symmetric case arises when the products are taken in pairs with opposite signs: that is

$$\phi^t(\mu) = \left(1+u^2/\lambda_1^2\right)^{\mu_1 t}\ldots\left(1-u^2/\lambda_n^2\right)^{\mu_n t}.$$

8.4.6. The Bessel process acquires its name because the form of the density involves a Bessel function, $I_t(x)$ (for $x > 0$):

$$f^t(x) = \left(\mathrm{e}^{-x}/x\right)tI_t(x), \quad \text{where } I_t(x) = \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k+t+1)}\left(\frac{x}{2}\right)^{2k+1}; \tag{8.36}$$

the characteristic function is given by

$$\phi^t(u) = \left\{1-\mathrm{i}u-\sqrt{\left[(1-\mathrm{i}u)^2-1\right]}\right\}^t. \tag{8.37}$$

The process deserves mentioning because it has the same interpretation as we put forward for the simple Poisson process, but referred instead to the Poisson variant of the Heads and Tails process (each occurrence of the phenomenon consists of a toss giving a gain of ± 1). Using the same notation, $t = T(y)$ and $y = T^{-1}(t)$ for the inverse function, we could say that the points at which y = an integer correspond to the instants

at which the gain $Y(t)$ first reaches level y (or that the 'integer part of $y = T^{-1}(t)$' is the maximum of $Y(t)$ in $[0, t]$, where $Y(t)$ is a Poisson Heads and Tails process).

8.4.7. The Pascal process is that for which $Y(t)$ has, for $t = 1$, a geometric distribution; for t = integer, a Pascal distribution; and, for general t, a negative binomial distribution (see Chapter 7, 7.4.4). Our previous discussion reveals that we are dealing with a compound Poisson process having a *logarithmic* distribution of jumps (see Chapter 6, 6.11.2) and having the positive integers as the set of possible values (with intensity $\mu = 1$).

The interpretation is similar to that of the two previous cases, except that here we do not have a density but, instead, concentrated masses (at the integer values). For t = integer, $Y(t)$ is the number of failures (in a Bernoulli process with probability p, where $\tilde{p}$ is the factor in the geometric distribution corresponding to $t = 1$) occurring before the tth success. In any interval of unit length (from t to $t + 1$), the increment has the geometric distribution; here, it can be thought of as generated by logarithmically distributed jumps, one per interval 'on average'. A noninteger t could be interpreted as 'the number of successes already obtained, plus the *elapsed fraction* of the next one'; $Y(t)$ would then be the number of failures that had actually occurred so far.

Of course, the remark made in the context of the gamma process (Section 8.4.5) applies equally to the other two processes (Sections 8.4.6 and 8.4.7).

8.5 Behaviour and Asymptotic Behaviour

8.5.1. We now turn to a probabilistic study of various aspects of the behaviour of the function $Y(t)$; as we pointed out in Section 8.1.6, these are, in fact, the most important questions. They might be concerned with the behaviour of $Y(t)$ in the neighbourhood of some given instant t (local properties), or in an interval $[t_1, t_2]$, or in the neighbourhood of $t = \infty$ (asymptotic properties). We might ask, for instance, whether $Y(t)$ vanishes somewhere in a given interval (and, if so, how many times), or if it remains bounded above by some given value M_1, or below by M_2 (or, more generally, by functions M_1(t) and $M_2(t)$ rather than constants), and so on. Asymptotically, we might ask whether these or other circumstances will occur from some point on, or locally, or only in the neighbourhood of some particular instant.

Phrases such as we have used here will have to be interpreted with due care, especially if one is not admitting the assumption of countable additivity. Particular attention must then be paid to expressing things always in terms of a finite number of instants (which could be taken arbitrarily large) and not in terms of an infinite number.

We have already come across a typical example of just this kind of question in Chapter 7, 7.5.3. This was the strong law of large numbers, which, in the case of a discrete process, consisted in the study of the asymptotic validity of inequalities of the form $C_1 \leqslant Y(t)/t \leqslant C_2$, or, equivalently,

$$C_1 t \leqslant Y(t) \leqslant C_2 t.$$

We shall see now how this problem, and generalizations of it, together with a number of other problems of a similar kind, can be better formulated, studied and appreciated by setting them within the more general context of the study of random processes. In the most straightforward and intensively studied cases, the processes that will turn out to

be useful are precisely those which we have been considering; specifically, the homogeneous processes with independent increments and, from time to time, the Wiener–Lévy process (in situations where it is valid as an asymptotic approximation).

It is both interesting and instructive to observe how the conjunction of the two different modes of presentation and approach serves to highlight conceptual links which are otherwise difficult to uncover, and to encourage the use of the most appropriate methods and techniques for each individual problem. In particular, later in this chapter (in which we shall only deal with the simplest cases) we shall see how the studies of the Heads and Tails process (developed directly using a combinatorial approach) and of the Wiener–Lévy process (which can be considered in a variety of ways) complement each other, enabling us in each case to use the more convenient form, or even to use both together.

8.5.2. We now consider two sets of questions, each set related to the other, and each collecting together, in a unified form, problems of different kinds which admit a variety of interpretations and applications, both theoretical and practical.

The first set brings together problems that reduce to the consideration of whether or not the path $y = Y(t)$ leaves some given region. In other words, whether or not it crosses some given barrier (the region may consist of a strip, $C_1 \leqslant y \leqslant C_2$, or $C_1(t) \leqslant y \leqslant C_2(t)$, or one of the bounds may not be present; i.e. $C_1 = -\infty$ or $C_2 = +\infty$). For this kind of problem, what we are usually interested in knowing is whether or not the process leaves the region and, if so, when this occurs for the first time, and if the process then comes to a halt. In the latter case, we have a so-called absorbing barrier. In general, however, it is useful to argue as if the process were to continue.

An analysis of this kind will serve to make the strong law of large numbers more precise, by examining the rate of convergence that can be expected (we shall, of course, have to make clear what we mean by this !). From the point of view of applications, there will be a number of possible interpretations, among which we note: the gambler's ruin problem (which could be thought of in the context of an insurance company); the termination of a sequential decision-making process because sufficient information has been acquired; the end of a random walk – for example, the motion of a particle – due to arrival at an absorbing barrier, and so on.

The second set of problems are all concerned with recurrence and involve the repetition of some given phenomenon (such as return to the origin, passing a check point etc.). We shall see that in this case the process divides rather naturally into segments and that it may be of interest to study various characteristics of these segments, which, in turn, often shed new light on the original recurrence problems.

In both cases, we shall first consider the Heads and Tails process and the Wiener–Lévy process, afterwards extending the study to the asymptotically normal cases. We shall also have occasion to consider other processes (the Bernoulli process with $p \neq \frac{1}{2}$, stable processes with $\alpha < 2$ and the Poisson process), mainly in order to have the opportunity of presenting further points of possible interest (explanations of unexpected behaviour, drawing attention to unexpected properties, and so on).

8.5.3. The three cases that we shall consider first involve the comparison of $|Y(t)|$ with $y = C$, $y = Ct$, $y = C\sqrt{t}$ ($c > 0$). Even though we shall dwell, in each case, on possible interpretations in an applied context, developing each topic as required, it will be useful

to bear in mind that these are in the way of preliminaries, enabling us subsequently to determine the foreseeable order of increase of $Y(t)$.

The first case, that of comparison with a constant, is the one of greatest practical interest and corresponds to the gambler's ruin problem under the assumption of fixed capital (i.e. with no increases or decreases, other than those caused by the game). We shall now consider the complex of problems which arise in this case, beginning with the simplest.

We first observe that the probability of $Y(t)$ lying between $\pm C$ tends to zero at least as fast as $K/\sqrt{t}$ (if $Y(t)$ has finite variance σ^2 per unit time, it tends to zero as $K/\sqrt{t}$, with $K = 2C/[\sqrt{(2\pi)}\sigma]$; if the process has infinite variance, then, whatever K may be, it tends faster than $K/\sqrt{t}$). The same holds if we are dealing with just a single t. If $\sigma = 1$, we have, in a numerical form, $K = 0{\cdot}8C$ (see Chapter 7, 7.6.3) and, for ease of exposition, we shall, as a rule, refer to this case (the bound is $0{\cdot}8C/\sqrt{t}$, and is approximately attained if $\sigma = 1$; if $\sigma = \infty$, we have strict inequality). The same holds for every interval $[a - C, a + C]$ of length $2C$.

To see this, note that, for $t = n$, in the case of Heads and Tails the maximum probability is given by $u_n \simeq 0{\cdot}8/\sqrt{n}$ and there are $2C$ integers lying between $\pm C$ (give or take $\pm$ 1[15]). In the normal case (Wiener–Lévy), the maximum density is $1/[\sqrt{(2\pi)}\sigma(t)]$, and $\sigma(t) = \sqrt{t}$. In the general case, with $\sigma < \infty$ (and, without loss of generality, we assume $\sigma = 1$), we have, asymptotically, the same process.

If $\sigma = \infty$, the bound corresponding to an arbitrary finite σ is *a fortiori* satisfied from a certain point onwards. Suppose $a > 0$ and arbitrary, but such that

$$\mathbf{P}\left(|X_h| < a\right) = p > 0,$$

further, let us distinguish the increments $X_h = Y(h) - Y(h - 1)$ according to whether they are $<a$ or $>a$ in absolute value; $Y(n) = \Sigma X_h$ then contains about np summands which are $<a$ (*truncated* distribution, with $\sigma < a$ finite), and the sum is already practically normal with density $<1/[\sqrt{(2\pi)}\sigma/t]$. Adding the sum of the other terms, we have, *a fortiori*, the same circumstance holding (theorem of the increase in dispersion; Chapter 6, 6.9.8).

What matters more than the quantitative result, is the qualitative conclusion: *however large C is chosen to be, the probability that* $|Y(t)|$ (*or* $|Y_n|$) *is greater than C differs from* 1 *by less than any given* $\varepsilon > 0$, *provided that t* (*respectively*, *n*) *is taken sufficiently large* (more precisely, from

$$t = n = \left(2/\pi\right)C^2/\varepsilon^2$$

onwards).

A fortiori, it follows that the probability of $|Y(\tau)| > C$ (or $Y_h > C$) for at least one $\tau \leqslant t$, or $h \leqslant n = t$, tends to 1 (more rapidly). In terms of the gambler's ruin problem, this implies that in a game composed of identical and independent trials between two gamblers each having finite initial capital, provided the game goes on long enough, the probability of it ending with the ruin of one of them tends to 1. Equivalently, the probability that the game does not come to an end is zero.

---

15 In order for this difference not to matter, it is, of course, necessary that C be much greater than 1. In general, in the case of lattice distributions, or distributions of similar kind, it is necessary that C be large enough for the concentrated masses to be regarded as a 'density'.

8.5.4. A warning against superstitious interpretations of the 'laws of large numbers'. *It is not only true that the absolute deviations, that is the gains and losses* (unlike the relative gains, the average gains per toss), *do not tend to offset one another and, in fact, tend to increase indefinitely in quadratic prevision, but also that it is 'practically certain that, for large, they will be large'* (and it is only in the light of the above considerations that we are now able to see this).

One should be careful, however, not to exaggerate the significance of this statement, turning it, too, into something misleading or superstitious. It holds for each individual instant t (or number of tosses n), but not for a number of them simultaneously. In fact, it does not exclude the possibility (and, indeed, we shall see that this is practically certain) of the process *returning to equilibrium* (and hence of there being segments in which $|Y(t)| < C$) every now and again (and although this happens more and more rarely, it is a never-ending process).

8.6 Ruin Problems; the Probability of Ruin; the Prevision of the Duration of the Game

8.6.1. We shall use p_h and q_n to denote the probabilities of ruin at the hth trial, or before the nth, respectively

$$\left(q_n = p_1 + p_2 + \ldots + p_n,\ p_h = q_h - q_{h-1}\right).$$

In the case of two gamblers, G_1 and G_2, we shall use p'_h and p''_h, q'_n and q''_n for the probabilities of ruin of G_1 and G_2, respectively

$$\left(p_h = p'_h + p''_h,\quad q_n = q'_n + q''_n\right),$$

and c' and c'' for their respective initial fortunes.

By q' and q'' (or q'_∞ and q''_∞), we shall denote the probabilities of ruin within an infinite time, to be interpreted as limits as $n \to \infty$: under the assumptions of Section 8.5.3, we have $q'_n + q''_n \to 1$, and therefore $q' + q'' = 1$.

The probability of ruin, in a fair game, is an immediate consequence of the fairness condition: the previsions of the gains of the two gamblers must balance; that is $q'c' = q''c''$, from which we deduce that $q' = K/c'$, $q'' = K/c''$ $(K = c'c''/(c' + c''))$. In other words, the probabilities of ruin are inversely proportional to the initial fortunes. More explicitly,

$$q' = c''/\left(c' + c''\right),\ q'' = c'/\left(c' + c''\right). \tag{8.38}$$

Comments. In these respects, we might also apply the term 'fair game' to a nonhomogeneous process with nonindependent increments, provided that the prevision conditional on any past behaviour is zero (such processes are called *martingales*). One could think, for example, of the game of Heads and Tails with the stakes depending in some way on the preceding outcomes. With these assumptions, any mode of participation in the game is always fair: it does not matter if one interrupts the play in order to alter the

stakes, or even if one decides to stop playing on account of a momentary impulse, or when something happens – such as someone's ruin.

The relation $q'c' = q''c''$ is an exact one if ruin is taken to mean the exact loss of the initial capital with no unpaid residue; in the latter case, it would be necessary to take this into account separately. If, for example, the jumps which are disadvantageous to G_1 and G_2 cannot exceed Δ' and Δ'', respectively, then c' must be replaced by some $c' + \theta'\Delta'$ ($0 \leqslant \theta' \leqslant 1$), with a similar substitution for c''; the error is negligible if the probable residues are small compared with the initial capital.

The conclusion holds exactly for the Wiener–Lévy process because the continuity of $Y(t)$ ensures that it cannot exceed c' and c'' by jumping past them. The same holds true for the Heads and Tails process – including the Poisson variant – provided c' and c'' are integers (because it is then not possible, with steps of ± 1, to jump over them).

It is clear, particularly if we use the alternative form

$$q' = 1 - 1/\left[1 + \left(c''/c'\right)\right],$$

that the probability of G_1's ruin tends to 1 if the opponent has a fortune that is always greater than his. If he plays against an opponent with infinite capital, the probability of ruin is, therefore, $q' = 1$ (and this is also true if one plays against the general public – who cannot be ruined). This is the *theorem of gambler's ruin* (for fair games).

The case of unfair games reduces to the previous case if one employs a device that goes back to De Moivre: in place of the process $Y(t)$, we consider $Z(t) = \exp[\lambda Y(t)]$. If λ is chosen in such a way as to make the prevision of $Z(t)$ constant (=1, say), then the process $Z(t)$ is fair, and ruin (starting from Z(0) = 1, which corresponds to $Y(0) = 0$) occurs when one goes down by $\overline{c}'' = 1 - \exp(-\lambda c')$, or up by

$$\overline{c}'' = \exp\left(\lambda c''\right) - 1.$$

The probabilities of ruin are, therefore, inversely proportional to $\overline{c}'$ and $\overline{c}''$. It only remains to say how λ is determined. We observe that $\exp[\lambda Y(t)] = \phi^t(-\mathrm{i}\lambda)$ and that, if it exists (see Chapter 6, 6.10.4), ϕ is real and concave on the imaginary axis, taking the value 1 (apart from at the origin) only at the point $u = -\mathrm{i}\lambda$, with λ positive if the game is unfavourable ($\mathbf{P}[Y(t)] < 0$).

Example. We consider the case of Heads and Tails with an unfair coin ($p \neq \frac{1}{2}$) and with gains ± 1. We have $\exp[\lambda Y(1)] = pe^{\lambda} + \tilde{p}e^{-\lambda} = 1$, in other words (putting $x = e^{\lambda}$), $px^2 - x + (1 - p) = 0$ for $x = 1$ and $x = \tilde{p}/p$; $x = e^{\lambda} = 1$ would yield $\lambda = 0$ (which is meaningless), so we take $e^{\lambda} = \tilde{p}/p$, $e^{-\lambda c'} = (\tilde{p}/\mathrm{p})^{-c'}$, $e^{\lambda c''} = (\tilde{p}/p)^{c''}$ from which we obtain

$$q' = \frac{\left(\tilde{p}/p\right)^{c''} - 1}{\left(\tilde{p}/p\right)^{c''} - \left(\tilde{p}/p\right)^{-c'}},\ q'' = \frac{1 - \left(\tilde{p}/p\right)^{-c'}}{\left(\tilde{p}/p\right)^{c''}\left(\tilde{p}/p\right)^{-c'}}. \tag{8.39}$$

If one plays against an infinitely rich opponent, the passage to the limit as $c'' \to \infty$ gives two different results, according to whether the game is favourable, $(\tilde{p}/p) < 1$, or unfavourable $(\tilde{p}/p) > 1$; in the latter case, $q' = 1$ (as was obvious *a fortiori*; ruin is practically certain in the fair case); if, instead, the game is favourable, the probability of ruin is

$q' = (\tilde{p}/p)^{c'}$ and $1 - q' = 1 - (\tilde{p}/p)^{c'}$ is the probability that the game continues indefinitely.

8.6.2. The prevision $\mathbf{P}(T)$ of the duration T of the game until ruin occurs can also be determined in an elementary fashion for the game of Heads and Tails (even for the unfair case) and then carried over to the Wiener–Lévy process.

Instead of merely determining $\mathbf{P}(T)$ (starting from $Y(0) = 0$), it is convenient to determine the prevision of the future duration for any possible initial value y ($-c' \leqslant y \leqslant c''$) using a recursive argument; we denote this general prevision by $\mathbf{P}_y(T)$. Obviously, we must have $\mathbf{P}_y(T) = 0$ at the end-points ($y = -c'$, $y = c''$) because there ruin has already occurred. For y in the interval between the end-points, we have, instead, the relation

$$\mathbf{P}_y(T) = 1 + \tfrac{1}{2}\left[\mathbf{P}_{y-1}(T) + \mathbf{P}_{y+1}(T)\right]$$

(because we can always make a first toss, and then the prevision of the remaining duration can be thought of as starting from either $y + 1$ or $y - 1$, each with probability $\frac{1}{2}$). We then obtain a parabolic form of behaviour (the second difference is constant!) with zeroes at the end-points; explicitly, we obtain

$$\mathbf{P}_y(T) = -(y + c')(y - c''), \qquad \text{and hence} \quad \mathbf{P}(T) = \mathbf{P}_0(T) = c'c''. \tag{8.40}$$

As c'' increases, $\mathbf{P}(T)$ tends to ∞ no matter what $c' > 0$ is; it follows that $\mathbf{P}(T) = \infty$ for a game against an infinitely rich opponent, so that although ruin is practically certain ($q' = 1$), the expected time before it occurs is infinite.

Even in the case where c' and c'' are finite, the expected duration of the game, although finite, is much longer than one might at first imagine. For example, in the symmetric case, $c' = c'' = c$, the expected duration of the game is c^2 tosses: 100 tosses if each gambler starts with 10 lire; 40 000 tosses if each starts with 200 lire; 25 million tosses if each starts with 5000 lire. In the most extremely asymmetric case, $c' = 1$, $c'' = c$, the expected duration is c tosses; 1000 tosses if initially the fortunes are 1 lira versus 1000 lire; 1 million if initially we have 1 lira versus 1 million. One should note, however, that in this asymmetric case the gambler whose initial fortune is 1 lira always has the same (high) probabilities of coming to grief almost immediately, whatever the initial capital of his opponent (be it finite or infinite), provided that it is sufficient to preclude the opponent's ruin within a few tosses. Specifically, the probability is 75% that the gambler with 1 lira is ruined within 10 tosses, 92% that he is ruined within 100 (in general, $1 - u_n \simeq 1 - 0{\cdot}8/\sqrt{n}$); in these cases, fortunes of 10 or 100 lire, respectively, will ensure that the opponent cannot be ruined within this initial sequence of tosses. On the other hand, there is a chance, albeit very small, that the opponent will be the one who is ruined (this is about $1/c$; one thousandth if $c = 1000$, for example). In order for this to happen, it is necessary that the gambler who begins with 1 lira reaches a situation of parity (500 versus 500) without being ruined; after this, the expected duration of the game will be $500^2 = 250\,000$ tosses, there then being equal probabilities of ruin for the two parties. There is, therefore, a probability of two in a thousand of reaching parity but, in such a case, the subsequent duration of the game is almost certainly very long. As always, one should remember that prevision is not prediction.

For the Wiener–Lévy process, thinking of it as a limit case of Heads and Tails, one sees immediately that exactly the same conclusion holds. It is sufficient to observe that the change of scale ($1/N$ for the stakes, $1/N^2$ for the intervals between tosses) leaves the duration unchanged: the initial capitals are Nc' and Nc'', the duration $N^2c'c''$, with $1/N^2$ as unit. More generally, we can say that, roughly speaking, the conclusion holds for all processes with finite variances ($\sigma = 1$ per unit time; otherwise, $\mathbf{P}(T) = c'c''/\sigma^2$) provided c' and c'' are large enough to make the ruin very unlikely after a few large jumps.

In the case of games which are not fair, one can apply the same argument, but the result is different. In the case of Heads and Tails with an unfair coin ($p \neq \frac{1}{2}$) the relation

$$\mathbf{P}_y(T) = 1 + (1-p)\mathbf{P}_{y-1}(T) + p\mathbf{P}_{y+1}(T)$$

reduces to the characteristic equation $py^2 - y + (1 - p) = 0$, with roots 1 and $(1 - p)/p$, which gives $A + B(\tilde{p}/p)^y$ as the solution of the homogeneous equation. It is easily seen that $y(1 - 2p)$ (or $y/(\tilde{p} - p)$) is a particular solution of the complete equation and, taking into account the fact that $\mathbf{P}_y(T) = 0$ for $y = -c'$ and $y = c''$, we have

$$\mathbf{P}_y(T) = \frac{1}{1-2p}\left[(y+c') - (c'+c'')\frac{1-(\tilde{p}/p)^{y+c'}}{1-(\tilde{p}/p)^{c'+c''}}\right]. \tag{8.41}$$

For the extension to the Wiener–Lévy process (and, more or less as we have said, to cases which approximate to it), it is sufficient to observe that, in the case we have studied, $m = 2p - 1$, $\sigma^2 = 1 - m^2$, from which we obtain $p = \frac{1}{2} + \frac{1}{2}m/\sqrt{(m^2 + \sigma^2)}$. Given the m and σ of a Wiener–Lévy process (or a general process), it suffices to evaluate p in this way.

If one plays against an infinitely rich opponent ($c'' = \infty$), we have $\mathbf{P}(T) = \infty$, provided the game is advantageous ($p > \frac{1}{2}$), and given that, with nonzero probability, it can last indefinitely. If it is disadvantageous ($p < \frac{1}{2}$) only the first term remains:

$$\mathbf{P}(T) = c'/(1-2p). \tag{8.42}$$

8.6.3. *The probabilities of ruin within given time periods* (i.e. within time t, or within $n = t$ tosses) provide the most detailed answer to the problem. Let us consider, for the time being, the case of one barrier ($c' = c$, $c'' = \infty$), and let us begin with Heads and Tails. We shall attempt to determine the probability, q_n, of ruin within n tosses; that is the probability that $Y_h = -c$ for at least one $h \leqslant n$.

The solution is obtained by making use of the celebrated, elegant *argument of Desiré André*. In the case of the game of Heads and Tails, we adopt the following procedure for counting the number of paths (out of the 2^n possible paths between 0 and n) which reach the level $y = -c$ at some stage. First of all, we consider those which terminate beyond that level, $Y(n) < -c$, and we note that there are exactly the same number for which $Y(n) > -c$, since any path in this latter category can be obtained in one and only one way from one of the paths in the former category by reflecting (in the straight line

$y = -c$) the final segment starting from the instant $t = k$ at which the level $y = -c$ is reached for the first time. In other words, we use the reflection

$$Y^*(t) = -c - \big(Y(t) + c\big) \qquad \big(k \leqslant t \leqslant n\big).$$

Finally, we note that (if $n - c$ is even) there are some paths for which $Y(n) = -c$. In terms of the probability (the number of paths divided by 2^n), the first group contribute to $\mathbf{P}(Y_n < -c)$ and, because of the symmetry revealed by André's reflection principle, so do the second group; the final group contribute to $\mathbf{P}(Y_n = -c)$. Expressing the result in terms of c rather than $-c$, we have therefore

$$q_n = 2\mathbf{P}\big(Y_n > c\big) + \mathbf{P}\big(Y_n = c\big) = \mathbf{P}\big(|\,Y_n\,| > c\big) \;+\; \tfrac{1}{2}\mathbf{P}\Big(\big|Y_n\big| = c\Big). \tag{8.43}$$

The basic idea is illustrated most clearly in Figure 8.6; to any path which having reached $y = -c$ finds itself at $t = n$ above that level, there corresponds, by symmetry, another path which terminates below it (and, indeed, if the first path terminates at $-c + d$, the second will terminate at $-c - d$; the reader should interpret this fact). Essentially, we could say that reflection corresponds to exchanging the rôles of Heads and Tails (from the instant of ruin onwards), a device that we have already come across (in example (*D*) of Chapter 7, 7.2.2).

A further principle – similar to that of Desiré André – was introduced by Feller (Vol. I, p. 20) under the heading of the duality principle; we prefer to call it the *reversal principle*, because the other term suggests connections, which do not actually exist, with other, unrelated, notions of 'duality'. The idea is that one reverses the time order; that is

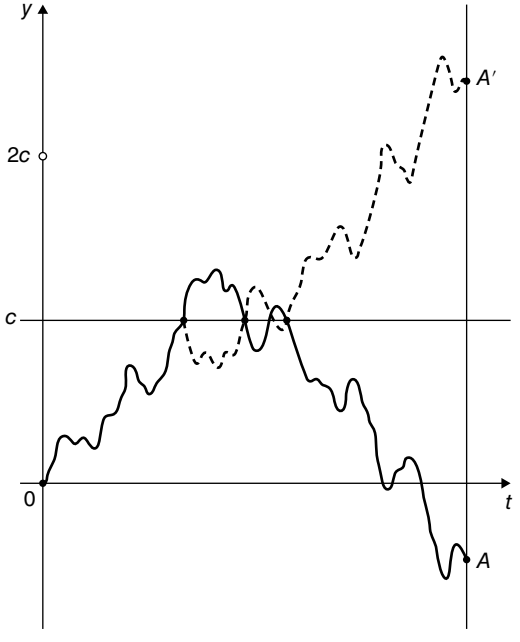


Figure 8.6 Desiré André's argument in the case of a single barrier. The paths which, after having reached level *c*, are below it at the end of the interval of interest (point *A*), correspond in a one-to-one manner, by symmetry, to those terminating at *A*′ (symmetric with respect to the barrier $y = c$). It follows (in a symmetric process) that the probability of ending up at *A*′, or at *A*, having reached level *c*, is the same.

The point 2*c* on the *y*-axis has been marked in because it corresponds to the 'cold source' in Lord Kelvin's method (Section 8.6.7).

the ordered events $E_1 \ldots E_n$ become $E_n \ldots E_1$, by setting $E_i = E_{n-i+1}$. The reversed gain is then

$$Y^*(h) = Y(n) - Y(n-h),$$

and the path is reversed (i.e. rotated by 180°) with respect to the central point $(\frac{1}{2}n, \frac{1}{2}Y(n))$.

The argument and the result given above have a far more general range of application. Indeed, the only facts about the distribution of the increments that we have used are its symmetry ($Y(t_0 + t) - Y(t_0)$ has equal probability of being $>a$ or $<-a$, and, in particular, of being $\lesseqgtr 0$), and the fact that the level $y = -c$ cannot be exceeded by 'skipping' it (i.e. anything exceeding it must actually pass through it). This holds for the Heads and Tails process if c = integer,[16] and for the Wiener–Lévy process (assuming it to be continuous) for arbitrary c. For other cases, it may have approximate, or asymptotic, validity (as in the *Remark* of Section 8.6.1), provided that the jumps in the direction in which the fixed level must be exceeded are small, or, at any rate, the large jumps are relatively rare (this brief comment will suffice; we do not wish to complicate matters by going into all the details).

Some further terminology and notation will be needed for certain more general problems and results that we wish to consider. The *maximum* of $Y(\tau)$ in $0 < \tau \leqslant t$[17] will be denoted by $\vee Y_n$ (an abbreviated form of $Y_1 \vee Y_2 \vee \ldots \vee Y_n$[18]) in the discrete case, and by $\wedge Y(t)$ in the continuous case; similarly, $\wedge Y_n$ and $\wedge Y(t)$ denote the *minimum*;

$$\left|\wedge Y(t)\right| = -\wedge Y(t) = \vee\left(-Y(t)\right)$$

is the absolute value of the minimum, and we shall refer to it as |minimum|.

With this notation, the q_n (or, to be more accurate, the $q_n(c)$) of equation 8.43 determine the probability distribution of $\vee Y_n$ (and of $|\wedge Y_n|$; it is the same, by symmetry);

$$q_n(c) = \mathbf{P}(\wedge Y_n \leqslant -c) = \mathbf{P}(\vee Y_n \geqslant c).$$

By subtraction, we obtain the probabilities

$$\begin{aligned}\mathbf{P}(|\wedge Y_n| = c) &= \mathbf{P}(\vee Y_n = c) = q_n(c) - q_n(c-1) \\ &= \mathbf{P}(Y_n = c) + \mathbf{P}(Y_n = c+1)\end{aligned} \tag{8.43′}$$

16 The set of levels which cannot be skipped may take various forms: either all c (positive, negative, or both) if $Y(t)$ varies continuously (nondecreasing, nonincreasing, or completely general); or all multiples of some given k (positive, negative, or both) if all positive jumps are $= k$, and the negative ones are multiples of k (or conversely, or if they are all $\pm k$); in all other cases there are no such levels.

17 We refer to $0 < \tau$ (instead of $0 \leqslant \tau$) in order to facilitate comparison with the discrete case (although the distinction loses its meaning in the continuous case); what is important is to stress that $\tau = t$ is to be included, and, indeed, we should stress that $Y(t)$ is to be understood as $Y(t + 0)$ (i.e. taking into account a possible jump occurring exactly at t).

18 The omission of Y_0 is irrelevant, except when it is useful to distinguish two cases that otherwise would both yield $\vee Y_n = 0$ (all $Y_n \leqslant 0$); with the convention adopted, we have, instead, $\vee Y_n = -1$, if $Y_1 = -1$, and the successive values are all $\leqslant -1$ (or, in general, stepping outside the example of Heads and Tails, $\vee Y_n$ can be an arbitrary negative value).

(only one of the two summands is ever present; the first if n and c have an even sum, the second if the sum is odd). Finally, we obtain

$$\mathbf{P}\left(\left|\wedge Y_n\right| = c\right) = \omega_h^{(n)} = \binom{n}{h} 2^{-n},$$

where either $2h - n = c$, or $2h - n = c + 1$.

It is important to pay particular attention to the cases $\wedge Y_n = 0$ and $\wedge Y_n = -1$. They are not contained in the general case (which is based on the assumption $c > 0$) but they can easily be reduced to the appropriate form. For $c = 1$, we have $\mathbf{P}(\vee Y_n < 1) = u_n$ (i.e. $\mathbf{P}(Y_n = 0)$ or $\frac{1}{2}\mathbf{P}(|Y_n| = 1)$, according to whether n is even or odd), and the two cases $\vee Y_n = 0$ and $\vee Y_n = -1$ are equally probable if n is even (and they do not differ by very much if n is odd). More precisely, we have $\mathbf{P}(\vee Y_n = -1) = \frac{1}{2}u_{n-1}$ (the first step is -1, and we do not then go up by $+1$), and, by subtraction

$$\mathbf{P}\left(\bigvee Y_n = 0\right) = u_n - \frac{1}{2}u_{n-1}$$

($u_n = u_{n-1}$ if n is even; otherwise $u_{n-1} = u_n n/(n-1)$). In words (for n even): u_n is also the probability that Y_h remains *non-negative* in $0 < t \leqslant n$ (and a similar argument holds for the *nonpositive* case).

Let us also draw attention to the following (interesting) interpretation of equation 8.43′.

The probability of attaining $y = c$ ($c > 0$) as the maximum level for $t \leqslant n$ is precisely the same as that of attaining the same level c (or $c + 1$, according to whether we are dealing with the even or odd case) at $t = n$ (but not, in general, as the maximum level).

To put it another way: the probability $2\omega_h^{(n)}$ that $|Y_n|$ assumes the value $c = 2h - n$ ($h > n/2$) splits into two halves for $\vee Y_n$; one half remains at c and the other at $c - 1$. If this partial shift of one is negligible in a given problem (as it is, in any case, asymptotically) we may say that the distributions of the absolute value, $\vee Y_n$, of the maximum, $\vee Y_n$, and of the |minimum|, $|\wedge Y_n|$, are all identical. Obviously, the maximum and the |minimum| are nondecreasing functions, and hence we can define their inverses. We denote by $T(y)$ ($y \geqslant 0$) the inverse of $\vee Y(t)$ and by $T(-y)$ the inverse of $\vee(-Y(t))$ (they also have the same probability distribution as processes; however, $T(y)$ for $y \lesseqgtr 0$ is not to be understood as a single process for $-\infty < y < +\infty$ but rather as a unified notation for two symmetric but distinct processes):

$T(y)$ = the minimum t for which

$$\vee Y(t) \geqslant y\,(y > 0), \quad \text{or} \quad \vee\left(-Y(t)\right) \geqslant -y\,(y < 0),$$

so that

$$\left(T(y) \leqslant t\right) = \left(\vee Y(t) \geqslant y\right)(y > 0) \vee \left(\vee\left(-Y(t)\right) \geqslant -y\right)(y < 0).$$

For every y, $T(y)$ is the random quantity expressing the instant (or, equivalently, the waiting time) either until ruin occurs, or until the first arrival at level (or point) y occurs, or until a particle is absorbed by a possible absorbing barrier placed at, and so on.

With the present notation, we can express, in a direct form, the probabilities of ruin (or of absorption etc.) as obtained in equation 8.43 on the basis of Desiré André's argument:

$$\begin{aligned}\mathbf{P}\left(\vee Y(t)\geq y\right)&=\mathbf{P}\left(\wedge Y(t)\leq -y\right)=\mathbf{P}\left(T(y)\leq t\right)=\mathbf{P}\left(T(-y)\leq t\right)\\&=2\mathbf{P}\left(Y(t)>c\right)+\mathbf{P}\left(Y(t)=c\right)\\&=\mathbf{P}\left(|Y(t)|>c\right)+\frac{1}{2}\mathbf{P}\left(|Y(t)|=c\right).\end{aligned}\tag{8.44}$$

We omit the term corresponding to $Y(t) = c$, which is only required in order to obtain exact expressions for the Heads and Tails case, and which is either zero or negligible in general (for the exact expressions of the Wiener–Lévy process, the asymptotic expressions for Heads and Tails, and for other cases). If we denote by $F^t(y)$ and $f^t(y)$ the distribution function and the density (if it exists) of $Y(t^{19})$, then the distribution function and the density of $|Y(t)|$ and therefore (either exactly or approximately) of $\vee Y(t)$ and $\vee(-Y(t))$ are given by

$$2F^t(y)-1 \quad (\text{for} \quad 0\leq y<\infty), \quad 2f^t(y) \quad (\text{for} \quad 0\leq y<\infty).\tag{8.45}$$

8.6.4. Substituting the exact forms for the Heads and Tails case into equations 8.44 or 8.45, we would have

$$q_n=\left(\tfrac{1}{2}\right)^{n-1}\sum_h\binom{n}{h}\left[0\leq h<\tfrac{1}{2}(n-c)\right]+\left(\tfrac{1}{2}\right)^n\binom{n}{(n-c)/2}\quad\left(\text{if } n-c \text{ is even}\right).\tag{8.46}$$

It is more interesting, however, to consider the approximation provided by the normal distribution; this will, of course, be exact for the Wiener–Lévy process and will hold asymptotically in the case of Heads and Tails, and for any other case with finite variance (which we shall always assume to be 1 per unit time). We have (for $y > 0$)

$$\begin{aligned}q_y(t)&=p_t(y)=\mathbf{P}\left(\vee Y(t)\geqslant y\right)=\mathbf{P}\left(T(y)\leqslant t\right)\\&=2\mathbf{P}\left(Y(t)\geqslant y\right)=\sqrt{(2/\pi)}\int_{y\sqrt{t}}^{\infty}\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x.\end{aligned}\tag{8.47}$$

[20]

Clearly, the form of equation 8.47, interpreted either as a function of y (with t as a parameter) or, alternatively, as a function of t (with y as a parameter), provides the distribution function of $|Y(t)|$ and $\vee Y(t)$ and that of $T(y)$ and $T(-y)$. It is often useful to have this expressed in terms of both the parameters y and t; it can then be interpreted

19 This would also be valid for processes other than those which are homogeneous with independent increments (satisfying the conditions stated for Desiré André's argument) where there is less justification for writing F^t and f^t. In fact, however, we shall not be dealing with the general case.

20 For y large (compared with $\sqrt{t}$), the approximation given by equation 7.20 (in Chapter 7, 7.5.4) can be used, and gives

$$q_y(t)=p_t(y)\simeq K\left(\sqrt{t}/y\right)\mathrm{e}^{-y^2/2t}, \qquad K=\sqrt{(2/\pi)}\simeq 0\cdot 8.\tag{8.47′}$$

according to whichever is appropriate. With this notation, we shall denote it in all cases by $\mathbf{P}(\vee Y(t) \geqslant y)$ (even in the Heads and Tails context).

The distribution of the maximum (or |minimum|) $\vee Y(t)$ (or $|\wedge Y(t)|$), and of $|Y(t)|$, is clearly the *semi-normal* (the normal distribution confined to the positive real axis), whose density is given by

$$f(x) = Kt^{-\frac{1}{2}} e^{-\frac{1}{2}x^2/t} \quad (x \geqslant 0), \qquad K = \sqrt{(2/\pi)} \simeq 0 \cdot 8. \tag{8.48}$$

This is half of the normal distribution with $\bar{m} = 0$ and $\bar{\sigma}^2 = t$; the mean and variance are given by $m = \sqrt{(2/\pi)}\bar{\sigma}$[21] and $\sigma = \sqrt{(1 - 2/\pi)}\bar{\sigma}$; that is, numerically, $m \simeq 0{\cdot}8\bar{\sigma}$ and $\sigma \simeq 0{\cdot}6\bar{\sigma}$ (these should not be confused!).

On the other hand, $T(y)$ (or $T(-y)$) has density

$$f(t) = Kyt^{-\frac{3}{2}} e^{-\frac{1}{2}y^2/t} \quad (t \geqslant 0), \qquad K = 1/\sqrt{(2\pi)} \simeq 0 \cdot 4. \tag{8.49}$$

This is the stable distribution with characteristic exponent $\alpha = \frac{1}{2}$ (which we mentioned in Section 8.4.4); in other words, it corresponds to jumps x whose density of intensity is $x^{-\frac{3}{2}}$. Since $\alpha < 1$, it has infinite prevision (in line with what we established directly in Section 8.6.2).

The fact that $T(y)$ had to have the stable distribution with $\alpha = \frac{1}{2}$ could have been deduced directly from the fact that

$$T(y_1 + y_2) = T(y_1) + \left[T(y_1 + y_2) - T(y_1)\right] = T(y_1) + T(y_2).$$

The time required in order to reach level $y_1 + y_2$ is, in fact, that required to reach y_1 plus that then required to proceed further to $y_1 + y_2$. However, given that, by the continuity of the Wiener–Lévy process, the level y_1 is reached (and not bypassed) at $T(y_1)$ with a jump, it is a question of going up by another y_2 under the same conditions as at the beginning. By virtue of the homogeneity, however, the distribution can only depend on y^2/t (and the density, in terms of y^2/t, could therefore only be – as, in fact, it actually is – a function of y^2/t divided by t).

We shall return (in Section 8.7.9) to the exact form for the Heads and Tails case, having encountered (for Ballot problems, in Section 8.7.1) an argument which enables us to establish it in a straightforward and meaningful way.

8.6.5. In the case where we consider *two gamblers*, G_1 and G_2 (with initial fortunes c' and c'', where $c' + c'' = c^*$), Desiré André's argument still applies, but now, of course, in a more complicated form. If we denote by A and B passages through levels c' and c'', respectively, a path whose successive passages are $ABABAB\ldots$ signifies the ruin of G_1; if the sequence begins with B, it signifies the ruin of G_2. (It does not matter how many

---

21 $m = \bar{\sigma}\left[2/\sqrt{(2\pi)}\right]\int_0^\infty x \exp\left(-\frac{1}{2}x^2\right) dx$;
it can easily be shown that the integral is equal to one.

times A is followed by B or B by A, nor does it matter whether the sequence ends in an A or a B; successive passages through the same level are not counted; e.g. $ABBAAABAABB = ABABAB$.) Desiré André's argument (as applied in the one-sided case) does not directly enable one to count the paths $\{A\}$ which signify G_1's ruin, nor the paths $\{B\}$ which signify G_2's ruin, nor the paths $\{0\}$ which indicate that neither is ruined (all belonging to the 2^n paths in the interval $[0, n]$). It does, however, enable one to count those of 'type (A)' 'type (B)' 'type (AB)', 'type (BA)', 'type (ABA)' and so on, where these refer to paths containing, in the sequence, the groups of letters indicated (which might be sandwiched between any number of letters).

Everything then reduces to the previous case of only one gambler; in other words, to

$$p_t(y) = p(y) = \mathbf{P}\left[\bigvee Y_n \geqslant y\right].$$

The probability of paths of 'type (A)' is, in fact, $p(c')$; that of paths of 'Type (AB)' is $p(c' + c^*)$ (because to first reach $-c'$ and then to reach $-c''$ requires a zigzag path along $c' + (c' + c'')$; this amounts to reflecting the path with respect to $y = -c'$, starting from the instant it reaches this level and continuing up until when it reaches c''); for 'type (ABA)' we have $p(c' + 2c^*)$, and so on. Similarly, for 'types' (B), (BA), (BAB), …, we have

$$p(c''), p(c'' + c^*), \qquad p(c'' + 2c^*), \ldots,$$

and in this way we arrive at the required conclusion.

The paths $\{A\}$ are, in fact, those given by

$$(A) - (BA) + (ABA) - (BABA) + (ABABA) - \ldots$$

(i.e. we start with those reaching $-c'$ and we exclude those first reaching c''; in this way, however, we exclude those that reach $-c'$ prior to c''; and so on). The same thing holds for $\{B\}$; the $\{0\}$ are those remaining (i.e. neither $\{A\}$ nor $\{B\}$) (see Figure 8.7).

It follows that the probabilities of ruin within n tosses are, for G_1, given by

$$q'_n = p(c') - p(c'' + c^*) + p(c' + 2c^*) - p(c'' + 3c^*) + \ldots \tag{8.50}$$

or (in terms of c''),

$$p(c^* - c'') - p(c^* + c'') + p(3c^* - c'') - p(3c^* + c'') + \ldots \tag{8.50′}$$

(there are a finite number of terms because $p(y) = 0$ when $y > n$).

The probabilities q''_n (of ruin for G_2) are obviously expressed by the same formulae, provided we interchange the rôles of c' and c″.

In particular, in the symmetric case, $c' = c'' = c$, we have

$$q'_n = q''_n = p(c) - p(3c) + p(5c) - p(7c) + \ldots \tag{8.51}$$

8.6.6. In the case of the Wiener–Lévy process (and, asymptotically, for Heads and Tails and for the asymptotically normal processes), we shall restrict ourselves, for simplicity and ease of exposition, to the symmetric case, which provides us with the

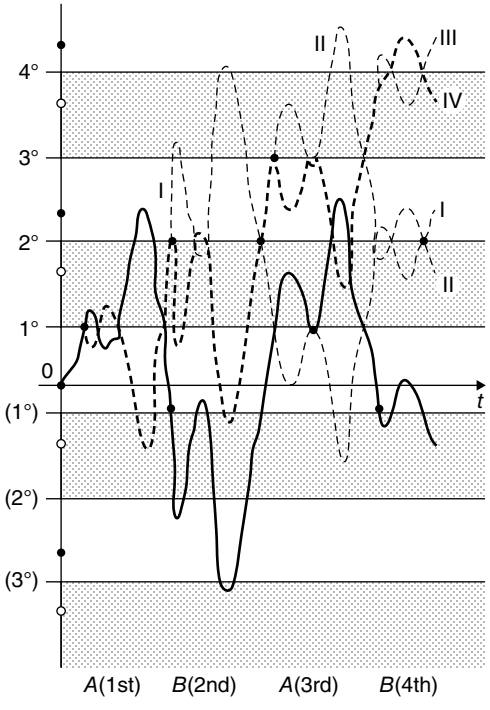


Figure 8.7 Desiré André's argument in the case of two barriers. The barriers are the straight lines bounding the white strip around the origin 0; the other strips are the *proper* images (white strips) and *reversed* images (dark strips) with the respective *hot* (black) and *cold* (white) sources (corresponding to Lord Kelvin's method; see Section 8.6.7). The actual path is indicated by the solid black line; its four successive crossings are denoted by *A*(1st), *B*(2nd), *A*(3rd), B(4th) (consecutive repeat crossings of *A* or *B* are not counted).

The final image of the path (obtained by repeated application of André's reflection principle) is indicated by the heavy broken line; it follows the same path up until *A*(1st) and is then given by the reflection (I) of it with respect to the 1st level. Then, after *B*(2nd), it is given by the reflection (II) of (I) with respect to the 2nd level, and so on. The continuations of the reflected paths (after the section in which they constitute the final image) are indicated by the lighter, dashed line. The image paths reaching the 1st, 2nd, 3rd levels, etc., correspond to paths of types *A*, *AB*, *ABA*, etc. (instant by instant); the same is true in the opposite direction (1st, 2nd, 3rd levels, etc. in the negative halfplane) for paths of types *BA*, *BAB*, etc.

distribution of $\vee|Y(t)|$, the maximum of the absolute values of Y in $[0, t]$. The ruin of one of the two gamblers within time t means, in fact, that, in the interval in question, Y reaches $\pm y$; that is that $|Y|$ reaches y.

In the case of Heads and Tails, we have

$$\mathbf{P}\left(\vee\left|Y\left(t\right)\right|\geqslant y\right)=q'_n+q''_n=2\sum_h\left(-1\right)^h p\left[\left(2h+1\right)y\right] \tag{8.52}$$

(there are, in fact, only a finite number of terms as we saw above). In the Wiener–Lévy case, $p(y)$ is given by equation 8.47 in Section 8.6.4, and hence

$$\mathbf{P}\left(\vee\left|Y\left(t\right)\right|\geqslant y\right)=2\sqrt{\left(2/\pi\right)}\sum_{h=0}^{\infty}\left(-1\right)^h\int_{(2h+1)y/\sqrt{t}}^{\infty}\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x. \tag{8.53}$$

Differentiating with respect to y, we obtain the density

$$\begin{aligned} f(y) &= K\sum_{h=0}^{\infty}(-1)^h(2h+1)\quad \mathrm{e}^{-\frac{1}{2}\left[(2h+1)y\right]^2/t} \\ &= K\left(\mathrm{e}^{-y^2/2t}-\mathrm{e}^{-(3y)^2/2t}+\mathrm{e}^{-(5y)^2/2t}-\mathrm{e}^{-(7y)^2/2t}+\ldots\right), \\ K &= 2\sqrt{(2/\pi t)}. \end{aligned} \tag{8.54}$$

8.6.7. It is instructive to compare the present considerations, based on Desiré André's argument, with those, essentially identical, based on Lord Kelvin's *method of images*, which is often applied to diffusion problems. We have already observed (Chapter 7, 7.6.5) that the Heads and Tails process can be seen, in a heuristic fashion, to approach a diffusion process, and that the analogy becomes an identity when we consider the passage to the limit which transforms the Heads and Tails process into the Wiener–Lévy process.

In order to use the method to formulate the problem of the ruin of *one gambler* (occurring when level $y = c$ is reached), it suffices to find the solution of the diffusion equation (given by equation 7.32 of Chapter 7, 7.6.5) in the region $y \leqslant c$ (where we assume $c > 0$), satisfying the initial condition of concentration at the origin, *together with the condition that it vanishes at* $y = c$. By virtue of the obvious symmetry (giving a physical equivalent to Desiré André's argument), it suffices to place initially, at the point $y = 2c$, a mass equal and opposite to that at the origin (the 'cold source'); this gives a process that is identical to the other, but opposite in sign, and symmetric with respect to the line $y = 2c$ instead of to $y = 0$. On the intermediate line, $y = c$, the two functions therefore cancel one another out and their sum provides the desired solution. The probability of ruin at any instant can be interpreted in terms of the flux of heat out past the barrier, together with an incoming cold flux; and so on.

In the case of two gamblers (i.e. barriers at $\pm c$, and the initial position of the mass at $y = 0$, or at $y = a$, $|a| < c$[22]), if we are to use the same trick we have to introduce an *infinite* number of sources, alternatively hot and cold – like alternate images of the face and the back of the head in a barber's shop with two mirrors on opposite walls. We have an infinite number of images of the mirrors (lines $y = (2k + 1)c$, k being an integer between $-\infty$ and $+\infty$) and in between them an infinite number of strips (proper and reversed images of the shop), and inside each of these strips the image ('hot' or 'cold') of the source. If the source is at the centre ($y = 0$), the others are at $y = 2kc$ (hot if k is even, cold if k is odd); otherwise – if it is at $y = a$ – the hot sources are at $2kc + a$ and the cold ones at $2kc - a$ (still with hot corresponding to k even, cold to k odd).

Using the techniques of this theory (Green's functions etc.) one can obtain solutions to even more complicated problems of this nature (e.g. those with curved barriers), where this kind of intuitive interpretation would necessitate one thinking in terms of something like a continuous distribution of hot and cold sources.

---

22 This is just a more convenient way of saying that one starts from 0 but places barriers at $c - a$ and $c + a$.

8.7 Ballot Problems; Returns to Equilibrium; Strings

8.7.1. We turn now to what we referred to in Section 8.5.2 as the second group of questions concerning random processes; those in which the study of the process reduces to an examination of certain segments into which it may be useful to subdivide it. More precisely, we shall consider the decomposition into *strings*; that is into the segments between successive returns to equilibrium (i.e. between successive *zeroes* of $Y(t)$).

Over and above their intrinsic interest, these questions often point the way to the formulation, understanding and solution of other problems that entail *recurrence* in some shape or form. In other words, problems relating to processes that, after every repetition of some given phenomenon (in this case, the return to equilibrium), start all over again, with the same initial conditions (or with modifications thereof which are easily taken into account).

The simplest scheme involving the notion of recurrence is that of events E_h forming a *recurrent sequence*[23] (such as the $E_h = (Y_h = 0)$ in our example) for which, when the outcomes of the preceding events are known, the probabilities depend on the number of events since the last success. In other words, the index (or 'time') $T^{(k)}$ of the kth success is the sum of the *k independent* waiting times $T_1, T_2, \ldots, T_k$: the T are integers in this case, but this is merely a special feature of this simple scheme.

We now provide a brief account of the most important aspects of the theory (for a fuller account, see Feller, Vol. I, Chapter XIII). We denote by f_h the probability that E_h is the first success (or, equivalently, that, following on from the last success obtained, the first success occurs in the hth place); in other words, $f_h = \mathbf{P}(T = h)$, where T = waiting time. It follows immediately that either $f = \sum f_h = 1$, or it is <1 (if the probability of a success does not tend to certainty as the number of trials increases indefinitely). We adopt the convention of denoting the difference $1 - f$ by f_∞ (the probability that the waiting time is infinite).

By convolution, we obtain the probabilities $f_h^{(2)}$ of E_h being the second success, and so on for $f_h^{(3)}$ and the rest. In general, we have

$$f_h^{(r)} = f_1 f_{h-1}^{(r-1)} + f_2 f_{h-2}^{(r-1)} + \ldots + f_{h-1} f_1^{(r-1)}. \tag{8.55}$$

Summing over r, we obtain the probability u_h of E_h being a success (without taking into account whether it was the first, second, …, or whatever; as above, this holds also for a success at the hth place following on from some success already obtained, assuming that nothing is known about successes for subsequent events):[24]

$$u_h = f_h + f_h^{(2)} + f_h^{(3)} + \ldots + f_h^{(h)} \qquad \left(\text{obviously, } f_h^{(r)} = 0 \text{ for } r > h\right). \tag{8.56}$$

---

23 These are usually called 'recurrent events', but this terminology does not fit in with ours (nor, in a certain sense, with the point of view we have adopted).

24 I would prefer to write $\mathbf{P}(E_h) = p_h$ (instead of u_h) as usual, in order not to make it appear that we are dealing with a special case, and in order to avoid confusion with the standard use of u_n (as the maximum probability for Heads and Tails). The reason we have used u_h is for the convenience of the reader who wishes to pursue this topic (which we are only scratching the surface of) using Feller: however, it only occurs in this section, so that there should be no cause for any confusion.

The sum of the $f_h^{(r)}$ gives f^r (=1 or <1, the same as f), and provides the probability that an rth repetition takes place. The sum

$$f + f^2 + f^3 + \ldots + f^r + \ldots$$

therefore gives the prevision of the total number of successes (finite or infinite, according to whether $f < 1$ or $f = 1$). The same prevision can be expressed in a different way, however, by the sum of the u_h; we therefore obtain

$$u = u_1 + u_2 + \ldots + u_h + \ldots = f/(1-f), \tag{8.57}$$

$$f = u/(u+1). \tag{8.57'}$$

If $f = 1$, $u = \infty$, the events E_h are called *persistent*; in the opposite case, $f < 1$, $u < \infty$, they are called *transient*. In the case of persistent events,[25] if we denote the prevision of the waiting time by τ, that is

$$\tau = f_1 + 2f_2 + 3f_3 + \ldots + hf_h + \ldots \qquad \left(\text{which could be } \infty\right), \tag{8.58}$$

then, as h increases, the probability u_h of success tends to the limit

$$\bar{u} = 1/\tau \qquad \left(\text{in particlar, } u_h \to \bar{u} = 0 \quad \text{if} \quad \tau = \infty\right). \tag{8.59}$$

Let us now return to the central topic of this section and explain how the 'Ballot problem' enters into the picture. This is simply a traditional way of referring to, and interpreting, a set of problems similar to those encountered under the heading of 'gambler's ruin', but relating now to drawings from an urn without replacement (which provides a model for the process of counting votes). The results also find application in statistics, where they form the basis of certain criteria (due to Kolmogorov and Smirnov) for considering the deviation of an empirical distribution function from a given hypothetical theoretical distribution. Anyway, although we shall refer to Ballot problems as a convenient aid to intuition, we shall think of this new case always in the context of Heads and Tails.

In fact, we shall study problems of Heads and Tails conditional on the knowledge of the number of successes, H, occurring in the N trials with which we are concerned (we have seen this argument before in Chapter 7, 7.4.3, where we used it to derive the hypergeometric distribution). Proceeding in this way, it is evident that, among other things, the graphical representation of this problem consists simply of the rectangular portion of the Heads and Tails lattice having opposite vertices at the origin (the starting point) and at $[N, H] = (N, 2H - N)$ (the point where the process terminates).

There are $\binom{N}{H}$ possible paths joining these two points (Figure 8.8); in order to fix ideas, we assume that $Y_N = 2H - N \geqslant 0$, that is $H \geqslant N - H$. For $0 < y < Y_n$, all paths either

25 The complication of 'periodic events' (E_h only possible for a multiple of some 'period' λ) can be avoided (and we assume this done) by confining attention to events $E_{h\lambda}$. Of course, this might not always be convenient in practice (e.g., if a sequence defined in terms of another sequence A_n, which is not periodic, turns out to be periodic: see Heads and Tails; returns to zero are only possible for n even).

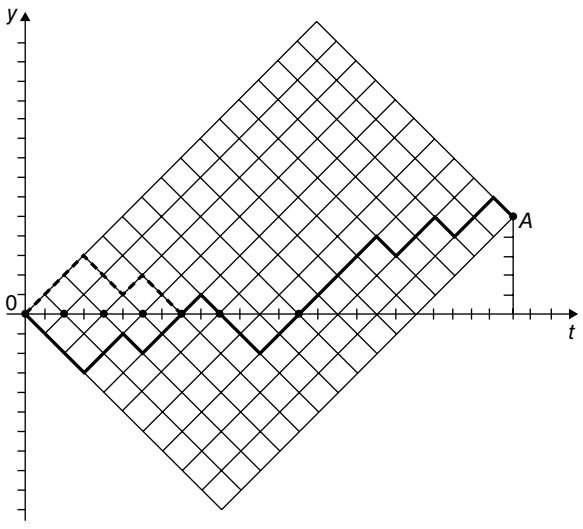


Figure 8.8 Desiré André's argument: the Ballot problem (i.e. the hypergeometric distribution). Paths from 0 to *A* which begin with a downward step correspond in a one-to-one fashion to those which start off upwards and then subsequently touch the *t*-axis (by symmetry in the interval before the *t*-axis is first reached).

cross or touch each barrier of the form y = constant. For the case $y = 0$ (and $y = Y_N$) we wish to consider how many paths touch it (either crossing it or not) after the initial $Y_0 = 0$ (or before the final $Y_N = N$, respectively). The same question (without the qualifications at the end-points) also arises for levels $y = Y_N + c$ ($c > 0$), or, equivalently,[26] for $y = -c$.

(*A*) The case $y = 0$, the Ballot problem, is the simplest one (we restrict ourselves to $Y_N > 0$ for now, postponing the case of $Y_N = 0$ to Section 8.7.3). All paths whose first step is downward must cross $y = 0$ again and, by reflecting the initial segment up to the first crossing, one obtains (with a one-to-one correspondence) all those having an upward first step which subsequently touch $y = 0$. But the first step has (like any other step) probability $(N - H)/N$ of being one of the $N - H$ downward steps; the probability of the eventual winner not being in the lead at some stage during the counting is therefore equal to twice this value, $2(N - H)/N$, and the probability that he is *always in the lead* during the counting is given by

$$1-2(N-H)/N=(2H-N)/N=Y_N/N. \tag{8.60}$$

(*B*) If we turn to the case $y = Y_N + c$ (or $y = -c$), $c > 0$, the principle of reflection (Desiré André) shows immediately that there are $\binom{N}{H+c}$ paths which touch this barrier in $0 \leqslant t \leqslant N$ (either crossing it or not, as the case may be). This is the number which finish up at the point whose ordinate is $(2H - N) + 2c = 2(H + c) - N$, the symmetric image of the given final point with respect to the barrier. In fact, these paths are obtained from the others (in a one-to-one onto fashion) by reflecting, with respect to the barrier, the final portion, starting from when the barrier is first reached (for $t = h$; $Y_h = Y_N + c$).

26 This is an obvious application of the reversal principle; see Section 8.6.3.

The probability s_c of reaching (and possibly going beyond) level $y = Y_N + c$ (or $y = -c$) is therefore given by

$$s_c = \binom{N}{H+c} / \binom{N}{H} = \frac{(N-H)(N-H-1)(N-H-2)\ldots(N-H-c+1)}{(H+1)(H+2)(H+3)\ldots(H+c)}. \tag{8.61}$$

The explicit expression is particularly useful when c is small (note that for $c = 1$ we have $s_1 = (N - H)/(H + 1)$), and is instructive in that it shows how the successive ratios $(N - H - c + 1)/(H + c)$ give the probability of reaching the required level $(Y_N + c)$ given that one knows that the level immediately below it (i.e. $Y_N + c - 1$) has been reached. The complementary probability is

$$(2H - N + 2c - 1)/(H + c),$$

and so the probability r_{c-1} that the *maximum* level reached is $Y_N + c - 1$ is given by the formula for s_c with $(2H - N + 2c - 1)$ replacing

$$(N - H - c + 1)$$

in the final factor; this is also the probability that the *minimum* level is $-(c - 1)$ (alternatively, one can obtain this by noting that $r_{c-1} = s_{c-1} - s_c$). Observe that $s_c = 0$ for $c \geqslant N - H + 1$ (Why is this so?).

(*C*) *The case of two barriers*. In the case of two barriers at levels $y = -c'$ and $y = Y_N + c''$ (c' and c'' positive), by performing successive reflections, as in the previous case, one obtains the paths terminating at the image points of the given final point ($Y_N = 2H - N$) with respect to the two barriers (thought of as parallel mirrors: there are an infinite number of images, but only those with ordinates lying between $\pm N$ can be reached). Setting $c^* = Y_n + c' + c''$, the distance between the barriers, the ordinates of the images are given by

$$(2k + 1)c^* - c' \pm c''.$$

(N.B.: for $k = 0$, we have $c^* - c' - c'' = Y_N = 2H - N$, the given final point, and

$$c^* - c' + c'' = Y_N + 2c'' = 2(H + c'') - N,$$

the unique image in the case of a single barrier $y = Y_N + c''$; in that case, we used c instead of c''.)

It follows that the probability of the lower barrier being reached first is given by

$$q'_N = \left(1/\binom{N}{H}\right)\left[\binom{N}{H+c'} - \binom{N}{H+c^*} + \binom{N}{H+2c^*+c'} - \binom{N}{H+3c^*} + \ldots\right] \tag{8.62}$$

(where we argue as in Section 8.6.5): the result for q''_N is similar (with c'' in place of c'). The sum, $q_N = q'_N + q''_N$, gives the probability of reaching a barrier (not distinguishing which, or which was reached first), and $1 - q_N$ that of not reaching either of them.

8.7.2. When studying a random process, it is often useful to consider it as subdivided into successive *strings*; that is into segments within which it retains the same sign. In our case,[27] this means those segments separated by successive *zeroes*, $Y_t = 0$, and necessarily having even length (since Y_t can only vanish for t even). Strings are either positive or negative (i.e. paths on the positive or negative half-plane, $Y_t > 0$ or $Y_t < 0$; see footnote 12 in Chapter 7, 7.3.2) and between any two strings the path has a zero at which it either *touches* the t-axis or *crosses* it, according to whether the two strings have equal or opposite signs.

If one thinks in terms of gains, that is of the excess of the number of successes over the number of failures, a zero represents a *return to equilibrium* (equal numbers of successes and failures; gain zero), and the string represents a period during which one of the players has a strict lead over the other. Omitting the word 'strict' and including the zeroes, we obtain periods in which one or other player does not lose the lead (i.e. the union of several consecutive strings having the same sign). One might also be interested in knowing the length of time, within some given period up to $t = N$, during which either player has had the lead. If one thinks of a random walk, the zero is a *return to the origin* and a string is a portion of the walk between two returns to the origin.

The above discussion, together with the results obtained so far, leads us directly into this kind of question, either with reference to the special case of Heads and Tails, or to that of the Ballot problem (which reduces to the former, if one thinks of $Y_N = 2H - N$ as being known).

8.7.3. *The Ballot problem in the case of parity:* $Y_N = 0$, *that is there are equal numbers,* $H = N - H = N/2$, *of votes for and against. What is the probability that one of the two candidates has been in the lead throughout the count*? In our new terminology, this means that the process forms a single string; that is there are no zeroes except at the end-points (if we are thinking in terms of a particular one of the candidates, the string must be of a given sign and the probability will be one half of that referred to in the question above).

This can easily be reduced to the form of case (*A*) considered in Section 8.7.1. In terms of the candidate who is leading before the last vote is counted, we must have $Y_{N-1} = 1$ (since we know that at the final step the lead disappears, and we end up with $Y_N = 0$); it follows that the required probability is given by

$$Y_{N-1} / (N-1) = 1 / (N-1)$$

This is the probability that one of the two candidates (no matter which) remains strictly ahead until the final vote is counted; the probability of this happening for a particular candidate is $1/2(N - 1)$.

8.7.4. *What is the probability that in a Heads and Tails process – or, more generally, in an arbitrary Bernoulli process – the first zero* (*return to equilibrium, passage through*

27 That of processes with jumps of ± 1, with paths on a lattice, and (for convenience) starting at the origin, $Y_0 = 0$. In other cases, one could have changes in sign without passages through zero occurring (and one could even have, in continuous time, a discontinuous process $Y(t)$ with an interval in which changes of sign occur within neighbourhoods of each point).

the origin) *occurs at time* $t = n$ (where n, of course, is even)? For this to happen, it is first of all necessary that $Y_n = 0$; the problem is then that considered in Section 8.7.3. The probability that if a zero occurs it is the first one is therefore equal to $1/(n-1)$. The probability of the first zero at $t = n$ is thus $\mathbf{P}(Y_n = 0)/(n-1)$. In other words, this is the probability that the first *string* (and hence any string, since the process can be thought of as starting all over again after every zero) has *length* n.

(*A*) In the case of Heads and Tails, we have $\mathbf{P}(Y_n = 0) = u_n$, so the probability of the first zero occurring at $t = n$ is given by

$$u_n / (n-1) \simeq \left(0{\cdot}8 / \sqrt{n}\right) / n = 0{\cdot}8 n^{-\frac{3}{2}}.$$

The probability that the string is of length n and has a given sign (i.e. is to the advantage of a particular one of two gamblers) is one half of this.

More precisely, we have (setting $n = 2m$)

$$\frac{u_n}{n-1} = \binom{2m}{m} / 2^{2m}(2m-1), \text{ and also } \frac{u_n}{n-1} = \frac{u_{n-2}}{n} = u_{n-2} - u_n.$$

Since

$$u_n \cdot 2^n = \binom{2m}{m} = \frac{2m!}{m!m!} = \frac{2m(2m-1)(2m-2)!}{m \cdot m \cdot (m-1)!(m-1)!}$$
$$= \frac{4(n-1)}{n}\binom{2m-2}{m-1} = \frac{4(n-1)}{n} u_{n-2} \cdot 2^{n-2},$$

we can verify at once that $u_n = u_{n-2}(n-1)/n$.

This establishes the following important conclusions:

(*a*) u_n is also the probability that there are no zeros up to and including $t = n$ (this is true for $u_2 = \frac{1}{2}$, and hence is true by induction, since $u_{n-2} - u_n$ is the probability of the first zero occurring at $t = n$);[28]
(*a*′) as in the footnote;

(*b*) since $u_n \to 0$, the probability that (as the process proceeds indefinitely) there is at least one return to equilibrium tends to 1 (and the same is therefore true for two, three or any arbitrary k returns to equilibrium);

(*c*) the form u_{n-2}/n tells us that $1/n$ is the probability that the string terminates at $t = n$ (since Y_n becomes 0), assuming that it did not terminate earlier (since u_{n-2} is the probability that $Y_t \neq 0$ for $t = 1, 2, \ldots, n-2$, and this is necessarily so for $t = n - 1 =$ odd);

28 It follows that the probability of Y_t $(0 < t \leqslant n)$ being *always positive* (or *always negative*) is $u_n/2$. If, instead, one requires only that (*a*′) is non-negative (or nonpositive), the probability is double: i.e. it is still u_n (as can be seen from Section 8.6.3; special cases of equation 8.43′ for $c = 0$ and $c = -1$).

(*d*) from this, we can deduce that u_n can be written in the form

$$u_n = \left(1-\frac{1}{2}\right)\left(1-\frac{1}{4}\right)\left(1-\frac{1}{6}\right)\left(1-\frac{1}{8}\right)\ldots\left(1-1/n\right);^{29} \tag{8.63}$$

in other words,

$$u_{n+2} = u_n\left(1-\frac{1}{n+2}\right) = \left(u_n\frac{n+1}{n+2}\right)$$

(as a product of complementary probabilities; a demographic analogy goes as follows: the probability of being alive at age n can be expressed as a product of the probabilities of not dying at each previous age);

(*e*) a further meaningful expression for u_n is given by

$$u_n = \sum_k \frac{u_k}{k-1}u_{n-k} \quad \left(\text{the sum being over } k, k \leqslant n\right); \tag{8.64}$$

observe that, in fact, each summand expresses the probability that the first zero is at $t = k$, and that there is another zero at $t = n$ (i.e. after time $n - k$); for $k = n$ (the final summand), we must take $u_0 = 1$, and hence $u_n/(n-1)$, a term which can be taken over to the left-hand side to give the explicit expression $u_n = [(n-1)/(n-2)]\,\Sigma'$ (where Σ' denotes the same sum, but without the final term).

We shall encounter further properties of u_n and $u_n/(n-1)$ later.

(*B*) In the general case (the Bernoulli process with $p \neq \frac{1}{2}$) we have instead

$$\mathbf{P}\left(Y_n = 0\right) = \binom{2m}{m}\left(p\tilde{p}\right)^m = u_n\left(4p\tilde{p}\right)^m = u_n\left[2\sqrt{\left(p\tilde{p}\right)}\right]^n,\ 2\sqrt{\left(p\tilde{p}\right)} < 1.$$

The probability of the first zero at $t = n$ is, therefore,

$$\left[u_n/\left(n-1\right)\right]\left[2\sqrt{\left(p\tilde{p}\right)}\right]^n < u_n/\left(n-1\right)\ \left(n \text{ even}\right), \tag{8.65}$$

and the sum of such probabilities is <1.

The remaining probability, P, given by $P(x) = 1 - \Sigma_n[u_n/(n-1)]\,(1-x)^n$ with $x - 1 - 4p\tilde{p}$, is the probability that a string has infinite length (which, with probability = 1, will be to the advantage of the favourite; i.e. the player with $p > \frac{1}{2}$). At the beginning of each new

---

29 We note that this enables us to establish Wallis's formula; from

$$u_{2m} = \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{5}{6}\cdot\frac{7}{8}\cdots\frac{2m-1}{2m} \simeq \frac{\sqrt{(2/2m)}}{\sqrt{\pi}},$$

we obtain

$$\sqrt{\frac{\pi}{2}} = \lim_{m\to\infty}\frac{2.4.6\ldots2m}{3.5.7\ldots(2m-1)}\cdot\frac{1}{\sqrt{(2m)}},$$

where π has its usual meaning, given by the double integral of Chapter 7, 7.6.7:

$$\int \mathrm{e}^{-\frac{1}{2}\rho^2}\,\mathrm{d}x\mathrm{d}y = \int \mathrm{e}^{-\frac{1}{2}\rho^2}\,\rho\,d\rho\,\mathrm{d}\theta = 2\pi.$$

string, there is probability $(1 - P)$ of a finite string, advantageous to one or other of the players, and probability P that the favourite embarks on an infinite string. The probability that the kth string turns out to be infinite is given by $P(1 - P)^{k-1}$.

Note, in particular, that property (b) only holds in the case of Heads and Tails (i.e. *only* if we have $p = \frac{1}{2}$). Otherwise, it is not at all *asymptotically certain* that a return to equilibrium takes place (and even less is it certain that such a return to equilibrium takes place an arbitrarily large number of times). On the contrary, it is asymptotically certain that the *favourite* (the player with $p > \frac{1}{2}$) maintains his lead from some given time onwards.

Remark. We have introduced the phrase '*asymptotically certain*' to mean that some given fact – for example, in the case under consideration, the occurrence of a return to equilibrium, or of k returns to equilibrium – has a probability tending to 1 of occurring in a random process, provided the process goes on indefinitely; that is if p_N is the probability that it occurs before time N, then $p_N \to 1$ as $N \to \infty$.

We note that if some given fact is asymptotically certain, then its occurrence k times (k arbitrary) is also asymptotically certain, provided (as in our case) that each time it occurs we find ourselves with the same initial conditions.[30] Without this latter stipulation, the conclusion of (b) – 'and the same is therefore true for two, …' – no longer holds (this is obvious but was not mentioned explicitly in (b) for the sake of brevity).

We note also that 'asymptotically certain' in no way (logically) implies 'certain' provided the process continues indefinitely (not even if we were to assume that we could examine the process in its entirety, placing ourselves beyond the end of time). It is even more important to note that the fact that the occurrence of an event k times (with k arbitrarily large) is asymptotically certain *does not imply* that, in a process of infinite duration, its occurrence *an infinite number of times* is certain (necessary), nor even that it has probability 1 (nor even that it is probable, or even possible). We can only say that this number of repetitions N (assuming that it makes sense to speak about it) is a random quantity (either integer or $+\infty$), which has probability 0 of taking on any individual finite value, and hence of belonging to any given finite subset of integers, such as those less than some preassigned integer k. It could, however, be certainly finite, like an 'integer chosen at random' (see Chapter 4, 4.18.3).

8.7.5. *What is the prevision of the length L of a string* (*i.e. of the waiting time t = L until the first zero*)? In the case of Heads and Tails, we see immediately that $\mathbf{P}(L)$ is infinite. In fact, $n(u_{n-2}/n) = u_{n-2}$; that is the contribution to the prevision corresponding to $L = n$ tends to zero like $n^{-\frac{1}{2}}$, and the sum diverges.

As we remarked above (see the discussion following equation 8.65), if $p \neq \frac{1}{2}$ this no longer happens (because of the presence of the factor $2\sqrt{(p\tilde{p})} < 1$ in the geometric progression): the (finite) strings have, in prevision, finite length. However, the prevision becomes infinite if we take into account the fact that each string could be the *final* one,

30 We are referring, therefore, to a recurrent sequence of events (see Section 8.7.1).

of *infinite*[31] length (but, if we distinguish between the two players, this only happens for the favourite: the first player if $p > \frac{1}{2}$).

Remark. The result that we are interested in (for Heads and Tails) is more conveniently expressed in the following form (which also gives us the opportunity to make an observation of a more general character). Each string consists, in prevision, of $\frac{1}{2}$ a string of length 2, of $\frac{1}{8}$ a string of length 4, ..., of $u_n/(n-1)$ a string of length n, and so on. In terms of the prevision of length, we have, in contrast, $2/2 = 1$ for strings of length 2, $4/8 = \frac{1}{2}$ for strings of length 4, ..., $n(u_{n-2}/n)$ for strings of length n and so on. Observe, in particular, that, for an individual string, the prevision of long strings is negligible (i.e. the prevision of strings $>n$ is less than any given $\varepsilon > 0$, provided n is taken sufficiently large), whereas, for the prevision of length (which is infinite), the prevision of the length of short strings (i.e. of those less than some preassigned, arbitrary, finite n) is negligible. It makes no difference if one makes the same statements but multiplies by 1000 (or a million, or whatever): 'out of 1000 strings, in prevision 500 have length 2, and their total length is, in prevision, 1000', and so on. Usually, one says 'on average'. We shall see later (Section 8.8.4) that there are dangers in using this form of expression.

8.7.6. *For the Ballot problem in the case of parity, what is the probability that one of the two candidates has never been behind during the count*? This is almost the same question as we asked in Section 8.7.3, except that we are here asking for something less: we admit the possibility that at some stages during the count the votes for the two candidates may also have been equal. It is sufficient that the lead of the candidate in question never becomes negative; that is that it never reaches the level $y = -1$. As in Section 8.7.3, we assume $Y_{N-1} = 1$; we are then back in case (B) of Section 8.7.1 and we can apply s_c for the case $c = 1$, which we have already given explicitly:

$$s_c = (N-H)/(H+1),$$

where we have to put $N-1$ in place of N and $N/2$ in place of H (since $Y_{n-1} = 2H - (N-1) = 1$). We hence obtain

$$(N-2)/(N+2) = 1 - 4/(N+2)$$

for the probability that level $y = -1$ is reached and $4/(N+2)$ is then the required probability. If we are thinking in terms of one particular candidate, the probability that the lead is never negative is one half of this, that is $2/(N+2)$.

Since the probability of the lead always being positive is $1/2(N-1)$ (case (A) in Section 8.7.1), we obtain, by subtraction, the probability that the lead is always non-negative, but sometimes zero: this probability is equal to the previous one multiplied by

$$3(N-2)/(N+2).$$

31 It seems out of place here to complicate such expressions in order to repeat critical comments of the kind mentioned in the previous 'Remark' (following (B)).

In a different form, assuming that the lead is always non-negative, the probability that it is always positive is $(N + 2)/4(N - 1)$, and the probability that it is zero is $3(N - 2)/4(N - 1)$ (i.e., for large N, they are practically equal to $\frac{1}{4}$ and $\frac{3}{4}$, respectively).

Remark. We have seen that $2/(N + 2)$ is the probability of a given candidate being ahead (whether strictly or not) either *always* or *never* (i.e. for N or 0 steps, paths either all in the positive or all in the negative half-plane). The possible values for the number of steps when he is in the lead are

$$0, 2, 4, \ldots, N-2, N,$$

and since there are $(N + 2)/2$ possible values, their average probability must be $2/(N + 2)$. But this is the probability in the two extreme cases we have considered and, moreover, by an obvious symmetry, the probabilities for h and $N - h$ are equal. It follows that either they are all equal, or they have a strange wavy behaviour – with at least three turning points. Actually, they are all equal: in other words, *we have a discrete uniform distribution over the steps spent in the lead.* The proof is not as straightforward as the statement and will be omitted,[32] in order not to interrupt the discussion and overcomplicate matters. A further justification for this is that the previous considerations have already made the conclusion highly plausible.

8.7.7. *What is the probability that in the Heads and Tails process* (*or, more generally, in any Bernoulli process*) *the first crossing of* $y = 0$ (i.e. the first zero where the path does not simply touch the axis) *occurs at time* $t = n$ (*where n, of course, is even*)? In other words, we are asking for the probability that the duration of the initial period during which one of the players *never falls behind* is equal to n; that is that this is the sum of the lengths of the initial consecutive strings whose sign is that of the first string. In order for this to happen, it is first of all necessary that $Y_n = 0$, and that no crossings have taken place for $t < n$; in addition, we require that the first toss after $t = n$ (i.e. the $(n + 1)$st) is opposite in sign to the very first toss (and thus to the strings already obtained). The probability we seek is therefore given by

$$\mathbf{P}\left(Y_n = 0\right).\left[4/\left(n+2\right)\right].2p\tilde{p},$$

where $4/(n + 2)$ is the probability of no crossing occurring (as determined in Section 8.7.6), and $2p\tilde{p}$ is the probability that the first and $(n + 1)$st tosses have opposite signs. In the case of Heads and Tails, $p = \tilde{p} = \frac{1}{2}$, this reduces to $2u_n/(n + 2)$ (or to $u_n/(n + 2)$ if one specifies which of the two players is to be ahead initially). In the general case $\left(p \neq \frac{1}{2}\right)$, the probabilities (for each finite n) are smaller and one has the residual probability of the lead being maintained indefinitely (by the favourite, as for the strings). Apart from a comment of this kind – made for comparative purposes – we shall restrict ourselves to the case of Heads and Tails.

---

32 We merely note that it follows from equation 8.64 of Section 8.7.4(*e*), and that the arguments are similar to those mentioned in Section 8.7.10 (for the arc sine distribution).

If we compare the result obtained for the length L of a string with that for the lead, V say, we have

$$\mathbf{P}(L=n)=u_{n-2}/n,\ \mathbf{P}(V=n)=2u_n/(n+2);$$

from this it is clear that

$$\mathbf{P}(V=n)=2\mathbf{P}(L=n+2),$$

and hence that

$$\mathbf{P}(V\geqslant n)=2\mathbf{P}(L\geqslant n+2)=u_{n+2}.$$

It is instructive to consider the implications of this; on the one hand for the first few values (small values, corresponding to short strings) and, on the other hand, asymptotically (large values, corresponding to long periods of lead).

In the case of the first few possible (even) values, we have:

$$\begin{array}{lccccc} n=2 & 4 & 6 & 8 & 10\ldots \\ u_n=\dfrac{1}{2} & \dfrac{3}{8} & \dfrac{5}{16} & \dfrac{35}{128} & \dfrac{63}{256}\ldots \end{array}$$

and hence

$$\mathbf{P}(L=n)=u_n/(n-1)=\frac{1}{2}\frac{1}{8}\frac{1}{16}\frac{5}{128}\frac{7}{256}\ldots$$

and

$$\mathbf{P}(V=n)=2u_n/(n+2)=\frac{1}{4}\frac{1}{8}\frac{5}{64}\frac{7}{128}\frac{21}{512}\ldots$$

As we know from the last remark, the values in the final line are twice those of the penultimate line each shifted one place to the left (except for the final one, which equals $\frac{1}{2}$; doubling the remaining values, whose sum is $\frac{1}{2}$, we obtain again the total probability 1). This direct comparison shows that for $n = 2$ the probability for L is greater (as is obvious: in order to have $V = 2$, we must have $L = 2$ and, moreover, the subsequent string must be of opposite sign). For $n = 4$, they are equal, and subsequently the probabilities for V become greater $\left(\frac{1}{16}=\frac{4}{64}<\frac{5}{64};\frac{5}{128}<\frac{7}{128};\frac{7}{256}=\frac{14}{512}<\frac{21}{512}\right)$. All this could be seen directly, by simply noting that the ratio $2(n-1)/(n+2)$ is equal to $2-6/(n+2)$.

For large values, this ratio is (asymptotically) equal to 2, and, in any case, $\mathbf{P}(V\geqslant n)\simeq 2\mathbf{P}(L\geqslant n)\simeq 2(0\cdot 8/\sqrt{n})=0\cdot 8/\sqrt{(n/4)}$. This can be expressed by saying that, in a sense, long periods in the lead are four times as long as long strings (more precisely, this is true in the sense that V has the same probability of reaching some (long) length n as L has of reaching length $n/4$).

8.7.8. *For the Ballot problem in the case of parity, what is the probability distribution of the maximum lead attained during the count by a particular candidate? What is the probability distribution of the absolute value of the lead? And why does it become conditional on the fact that a given candidate never lost the lead throughout the count? Or was strictly in the lead throughout the count*? Clearly, if we drop the reference to the Ballot problem, we see that we are dealing with the most general question of the probability distribution of $\vee Y_N$, or of $V|Y_N|$ (either for Heads and Tails, or for any Bernoulli process), assuming $Y_N = 0$, and possibly, also, $Y_t \geqslant 0$, or even $Y_t > 0$, for $0 < t < N$. This last assumption is the most restrictive and it amounts to seeking the probability distribution of the maximum *in a single string*. Under the next to last assumption, we could be dealing with a segment composed *either of a single string or of several consecutive strings all having the same sign*. In the general case, on the other hand, the segment might consist of a single string or of several, with arbitrary signs: it is only in this latter case that we need to distinguish between $\vee Y_N$ and $V|Y_N|$.

In fact, these are simply variants of problems (*A*) and (*B*) in Section 8.7.1. We shall consider them separately.

(*a*) The only assumption is that $Y_N = 0$ and we seek the distribution of the maximum of Y_t. From (*B*) of Section 8.7.1, we see, taking $H = N/2$, that $s_0 = 1$, and, for $c \geqslant 1$,

$$s_c = \mathbf{P}\left(\vee Y_N \geqslant c\right) = \frac{N(N-2)(N-4)\ldots(N-2c+2)}{(N+2)(N+4)(N+6)\ldots(N+2c)} \qquad \left(s_c = 0 \text{ for } c \geqslant (N+2)/2\right), \tag{8.66}$$

$$t_c = \mathbf{P}\left(\vee Y_N = c\right) = s_c - s_{c+1} = \frac{4c+2}{N+2c+2} s_c \tag{8.67}$$

(in particular, $r_0 = 2/(N + 2)$, as we already know). Applying Stirling's formula as given in equation 7.30 of Chapter 7, 7.6.4, we have, approximately (for c large, but $2c/N$ small, i.e. N much larger still), $s_c \simeq e^{-2c^2/N}$, $r_c \left(4c/N\right)\mathrm{e}^{-2c^2/N}$.

(*b*) Continuing with $Y_N = 0$ as the only assumption, we seek the distribution of the maximum of $|Y_t|$. Arguing as in (*C*) of Section 8.7.1 but also taking into account the symmetry ($N = 2H$; i.e. $Y_N = 0$, $c' = c'' = c$, $c^* = 2c$), we find that the probability $\overline{s}_c$ of either reaching or crossing $\pm c$ (which was denoted in (*C*) by $q_N = q'_N + q''_N$, here $q'_N = q''_N$) is given by

$$\overline{s}_c = \left(2 / \binom{2H}{H}\right)\left[\binom{2H}{H+c} - \binom{2H}{H+2c} + \binom{2H}{H+3c} - \binom{2H}{H+4c} + \ldots\right]. \tag{8.68}$$

Expressed more simply, using the s_c from (*a*) above, we have

$$\overline{s}_c = s_c - s_{3c} + s_{5c} - s_{7c} + \ldots, \tag{8.68′}$$

and similarly

$$\overline{r}_c = r_c - r_{3c} + r_{5c} - r_{7c} + \ldots.$$

Asymptotically, we therefore see from the previous expression that

$$\overline{s}_c = 2\sum_{k=1}^{\infty}(-1)^{k+1}\,\mathrm{e}^{-2k^2c^2/N},\ \overline{r}_c = (8c/N)\sum_{k=1}^{\infty}(-1)^{k+1}\,k\mathrm{e}^{-2k^2c^2/N}. \tag{8.68''}$$

(*c*) Let us now assume, in addition to $Y_N = 0$, that Y_t has not changed sign throughout $0 \leqslant t \leqslant N$ and, in order to fix ideas, let us assume that it is non-negative. It therefore makes no difference whether we talk in terms of the maximum of $|Y_t|$ or of Y_t (or of $-Y_t$, had we made the opposite assumption). We again argue as in (*C*) of Section 8.7.1, but now with $c' = 1$, $c'' = c$, in order to obtain the probability that Y_t always remains strictly between -1 and c. Dividing this by $2/(N + 2)$ (the probability that $0 \leqslant Y_t$, i.e. that $-1 < Y_t$), we obtain the probability of $\vee Y_t < c$ conditional on the given hypothesis; that is the $1-\overline{s}_c$ of the present case. If we use s_c to denote the same probability as in case (*a*), we have, therefore,

$$\begin{aligned}\overline{s}_c &= 1-\frac{N+2}{2}\\ &\quad\times\left(s_1+s_c-s_{c+2}-s_{2c+1}+s_{2c+3}+s_{3c+2}-s_{3c+4}-s_{4c+3}+\ldots\right),\text{ etc.}\end{aligned} \tag{8.69}$$

(*d*) This proceeds similarly: under the assumption of having been strictly in the lead ($Y_t \geqslant 1$, $0 < t < N$), we can reduce to the previous case by taking the axis $y = 1$ as the base line on the interval from $t = 1$ *to* $t = N - 1$ ($Y_1 = Y_{N-1} = 1$); for large N, the difference is very small.

8.7.9. *Similar problems for an arbitrary segment of the Heads and Tails process* (i.e. for a segment $0 \leqslant t \leqslant n$, where we do not assume, as in the previous examples, that $Y_N = 0$). The segment could consist of one string, or several strings, or none, and, in general, it will end in an incomplete string. We shall give a brief review of certain problems and their solutions, in order to draw attention to various of the points which need considering.

(*a*) So far as *periods in the lead* are concerned (see Section 8.7.6 and the *Remark*), we know, from Section 8.6.3, that u_n is the probability of Y_t remaining non-negative (in $0 \leqslant t \leqslant n$); that is that the lead is maintained for n steps out of the n (and the same holds true, obviously, for 0 out of n). Assuming n to be even, the number of steps spent in the lead can be any of

$$0,\ 2,\ 4\ldots,\ n-2,\ n.$$

We therefore have $(n + 2)/2$ possible values and the average probability is $2/(n + 2)$. However, the extreme cases have probabilities $u_n > 2/(n + 2)$,[33] so it is likely (by the same kind of argument that the *Remark* of Section 8.7.6 led us to believe that in that case, a segment consisting of complete strings, the probabilities were equal) that, in the general

33 Since it is the maximum of $n + 1$ probabilities $\omega_h^{(n)}\ (h = 0,\ldots,n)$, u_n is certainly $>1/(n + 1)$. It is, in fact, much greater than this, becoming more and more so as n gets larger: asymptotically, $u_n \simeq (2/(n+2)).0.4\sqrt{n}$.

case, there is a very small probability that a subdivision into periods of lead will consist of nearly equal lengths, and much greater probabilities for the less equal subdivisions. There is another consideration that makes this plausible. We already know that for the segment leading up to the last zero we have equal probabilities for all subdivisions, and that, from the last zero on, the lead does not change hands. In any case, the fact is that this turns out to be true and, in precise terms, we obtain $p_h = u_n u_{n-h}$ (h and n even) for the probability of being in the lead for h out of the n steps. The proof is much more difficult than one might expect from the simplicity of this formula, and we shall omit it.[34] We shall restrict ourselves to a consideration of how this probability behaves.

Recalling that $u_{h+2}/u_h = (h + 1)/(h + 2)$, we see that the ratio

$$p_{h+2}/p_h = (h+1)(n-h)/(h+2)(n-h-1)$$

is less than, equal to, or greater than 1, according to whether $h + 1 \lessgtr n/2$: taking the asymptotic expression for u_n, we have $p_h = u_h u_{n-h} = (2/\pi)\sqrt{/}[h(n - h)]$. In the limit, we can say that the proportion of time during which a gambler is ahead, in a long period of play, is a random quantity X, whose probability distribution has density $f(x) = 1/(\pi\sqrt{/}[x(1 - x)])$. This is the '*arc sine*' distribution, so-called because the distribution function, $F(x) = \int f(x)\mathrm{d}x$, is equal to $(2/\pi)\sin^{-1}(\sqrt{x})$, which might be better written as

$$(1/\pi)\cos^{-1}(2x-1).$$

We shall come back to this again and again.

(*b*) We know that the *probability distribution of* Y_n is the *Bernoulli* (or binomial) distribution ($p_h = \omega_h^{(n)}, h = 0,1,\ldots,n$), provided we know only that $Y_0 = 0$ (this holds similarly if we are given certain values, of which the last one is $Y_k = y$ with $k < n$; we then have $p_{y+h} = \omega_h^{(n-k)}$). The distribution is the *hypergeometric* if, in addition to Y_0, we are also given a value Y_N, $N > n$ (and similarly if two arbitrary values are known, one before and one after; $Y_{t'}$ and $Y_{t''}$, say, with $t' < n < t''$).

In general, we can say that any change in the state of information produces a change in the probabilities. In particular, if, in addition to knowing the initial value $Y_0 = 0$ (and, possibly, a subsequent value, or two arbitrary values, one before and one after), one also knows that Y_t has remained non-negative throughout $0 \leqslant t < n$, we have a range of possibilities as discussed above. The same holds if we have non-negativity throughout the entire process, $0 \leqslant t \leqslant N$, or throughout $t' \leqslant t \leqslant t''$ in the case where we know the values at the two points on either side, or even only for $t' \leqslant t < n$, or $n < t \leqslant t''$. Different probabilities also result if we assume the process to be strictly positive or strictly negative, or above or below some given level, or in between two levels, and so on. All this is obvious, but it needs emphasizing, and should be borne in mind.

(*c*) *The probability that level* $y = c$ *is reached for the first time at* $t = n$ ($Y_n = c, \wedge Y_{n-1} < c, c > 0$; symmetrically if $c < 0$) *is equal* (by the principle of reversal the problem is unchanged) *to the probability that* $Y_n = c$ *without there being any zeroes beforehand* (that is all Y_t, $0 < t < n$, have the same sign as $c = Y_n$). This probability is given by $\mathbf{P}(Y_n = c)$ multiplied by c/n, where c/n is *the probability that*, the value c of Y_t at $t = n$ being known,

34 See the notes in Section 8.7.10.

this level c has been reached for the first time at that point, and also the probability that the starting level ($y = 0$) has not been reached again; that is that Y_t *has always had the same sign as c*.

If we denote by H one or other of the two hypotheses mentioned, we have that $\mathbf{P}(H) = u_{n-1}/2$ (Section 8.6.3) and that $\mathbf{P}(H|Y_n = c) = c/n$ (the Ballot problem: case (A) of Section 8.7.1). The probability that we seek (which can also be obtained as the probability of ruin occurring precisely at the nth toss, $p_n(c) = q_n(c) - q_{n-1}(c)$; see Sections 8.6.1 and 8.6.3) therefore has the value stated:

$$\mathbf{P}\left[\left(Y_n = c\right).H\right] = \left(c/n\right)\mathbf{P}\left(Y_n = c\right) = \left(c/n\right)\omega_h^{(n)} \quad \left(c = 2h - n, c > 0\right). \tag{8.70}$$

The probability distribution of Y_n conditional on H (one of the two hypotheses) is, therefore, proportional to $c\omega_h^{(n)}$; in fact, we have

$$\begin{aligned}\mathbf{P}\left(Y_n = c\middle|H\right) = \mathbf{P}\left[\left(Y_n = c\right).H\right] / \mathbf{P}\left(H\right) = \left(2 / nu_{n-1}\right)c\omega_h^{(n)} \\ \left(c = 2h - n > 0\right).\end{aligned} \tag{8.71}$$

Expressed in words: the probabilities of the possible values c for Y_n are altered (by the condition H) in a manner proportional to c (those for $c \leqslant 0$ are clearly 0), and the normalization factor K has been found explicitly in passing. The same result (and proof) goes through for the opposite hypothesis and provides the probability that $Y_n = c$ given that Y_n is greater than any value obtained previously (for $0 \leqslant t < n$, all Y_t are $< Y_n$; in this case, of course, we do not exclude the possibility of negative values). Asymptotically, we have a distribution of the form $f(x) = Kx\mathrm{e}^{-x^2/2} (x \geqslant 0)$.

(*d*) We now restrict ourselves to the special case of knowledge of one value before and one after, $Y_0 = 0$ and $Y_N = 0$, together with the condition that $Y_t > 0$ throughout the given interval. The distribution of Y_n (for any integer n, $0 < n < N$) is obtained in a similar way, by observing that the probability that $Y_n = c$ ($c = 2h - n > 0$), and that the given conditions hold, is, setting $N = 2H$, given by

$$\left[\binom{H}{h}\binom{H}{n-h} / \binom{2H}{n}\right].\left(c/n\right).\left(c/\left(N-n\right)\right)\Big] \tag{8.72}$$

(the product formed by taking $\mathbf{P}(Y_n = c)$ from the hypergeometric case and multiplying it by the probability that the process does not vanish in the passage from 0 to c during the first n and subsequent $N - n$ steps). We note, however, that this probability can also be thought of as the product of the p_h we are after, multiplied by the probability of the hypothesis that at time $t = N$ we obtain the first zero, the sign having previously been positive. This latter probability is equal to $u_n/2$, so we obtain

$$p_h = \left[\frac{2}{u_n n\left(N-n\right)\binom{N}{n}}\right] c^2 \binom{H}{h}\binom{H}{n-h} = K.c^2.\bar{\omega}_h^{(n)} \tag{8.73}$$

(we have placed a bar over the ω in order to stress the fact that it is the ω of the hypergeometric rather than the Bernoulli process, as above). One should note, however, the meaningful analogy between the two (every condition of positivity, on the left and on

the right, entails a modification proportional to c). Asymptotically, we have a distribution of the form $f(x) = Kx^2 e^{-x^2/2}\,(x > 0)$.

8.7.10. *Remark*. It is instructive to consider in more detail some of the problems which lead (asymptotically) to the arc sine distribution (see Section 8.7.6 and the *Remark* in Section 8.7.9 (a)). We have omitted the proofs and we suggest that the reader refers to the third edition of Feller, Vol. I (1968). Comparison with earlier editions will reveal the simplifications in formulation that took place from one edition to the other, partly as a result of greater insight into the problems and their connections with one another, and partly for purely fortuitous reasons (see Chapter III, Section 4 of Feller, 'Last visit and long leads', and, in particular, the historical notes on pages 78 and 82).

The arc sine distribution was considered by P. Lévy (1939) in connection with the Wiener–Lévy process (see Section 8.9.8).[35] The application to Heads and Tails and other cases (obvious in the asymptotic case) seemed to require 'mysterious' forms of explanation, until their combinatorial character was revealed (by Sparre Andersen, in 1953). The methods used were quite complicated, and still were in the first edition of Feller (where they were due to Chung and Feller); the new feature, which arises in the passage from the second to the third edition, lies in the preliminary statement of a simpler theorem, which, from a qualitative viewpoint, begins to explain the weighting towards very unequal subdivisions of periods in which the lead does not change hands.

We can prove this in just a few lines. *The probability that in* $2m$ *tosses at Heads and Tails the final return to zero,* $Y_t = 0$, *occurs at* $t = 2k$ *is given by* $u_{2k}u_{2m-2k}$ (which is the discrete version of the arc sine distribution). In fact, the probability that $Y_{2k} = 0$ is u_{2k}, and the probability of no zeroes in the $2m - 2k$ subsequent tosses is u_{2m-2k} (see Section 8.7.4(a)). This is trivial, and yet the theorem is new (according to Feller); moreover, it was discovered by chance, experimentally, on the basis of observed statistical properties of random sequences produced by a computer. These were detected by capable mathematical statisticians, who then simply pointed out, and subsequently proved, that the distribution was symmetric (without realizing that it was the arc sine distribution).

This is by no means intended as in any way disrespectful to the number of authors who have made valuable contributions to this topic. It merely goes to show that tucked away in the vast rock-pile of problems there is the odd nugget lying unobserved; once noticed, of course, it appears obvious.

The following little calculation, which I made simply out of curiosity, may be new and possibly of some interest. I observed that it was not appropriate to call one gambler 'luckier' simply because he has led for most of the game so far (see the footnote to

35 I (vaguely) remember that an obvious property of the arc sine distribution – the density taking its maxima at the end-points – was considered paradoxical, even in cases where it was natural, as for observations of periodic phenomena (for example, in the case of a river flooding, the level remains around the maximum longer than it does around intermediate values, which are passed through more rapidly, both when the level is increasing and decreasing): see Figure 8.9. Of course, when periodicity is crude (for example, seasonal temperature changes, where maxima and minima vary from year to year) there are smoothed peaks, or sequences of peaks.

Section 8.8.1): it might well turn out that his luck runs out at the end of the game – 'he who laughs last laughs longest'. The probability of this happening is given by

$$1/\sqrt{3\pi} = 0\cdot 184 = \int_{\frac{1}{2}}^{1} \left(2\pi t\sqrt{\left[t(1-t)\right]}\right)^{-1} \mathrm{d}t. \tag{8.74}$$

To see this, let t denote the time when the final zero occurs (taking [0, 1] as the whole interval). If $t < \frac{1}{2}$, the one who is leading at the end is also the one who has spent most time in the lead. If $t > \frac{1}{2}$ in order for the one who is leading at the end to have spent most time in the lead, he must previously (i.e. before time t) have spent at least an additional time $t - \frac{1}{2}$ in the lead. Because of the uniformity of the distribution of the lengths in an interval between two zeroes, this, for given t, has probability $\left(t-\frac{1}{2}\right)/t = 1-\frac{1}{2}t$. This leaves a probability $\frac{1}{2}t$ for the opponent, conditional on t (having the arc sine distribution) being greater than $\frac{1}{2}$; we thus obtain equation 8.74)

8.8 The Clarification of Some So-Called Paradoxes

8.8.1. We have already found (and will do so again) that certain of the conclusions we have arrived at have had a paradoxical air, or, at any rate, have been easy to misinterpret. We have discussed various examples where such misinterpretation arises and, in so doing, have attempted to clarify the issues involved. In particular, we recall the laws of large numbers and, in the gambling context, the long expected time to ruin. The topics we have just been considering also lend themselves to discussions of this kind. Indeed, it is hard to decide whether their main value lies in the knowledge they provide, and the light they throw on a number of important theoretical and practical questions, or in the opportunity they give one to clear up a number of misconceptions and confusions, which otherwise could make one rather wary of entering into the probabilistic domain at all.

In those aspects of the Heads and Tails process that we have just been studying, it surely seems rather strange and mystifying that some kind of 'stationarity' or regularity does not hold. In particular, why is there not a tendency for the periods of unbroken lead to be equally distributed in the two opposite directions (all the more so after having seen that the process can be considered as an indefinite sequence of *strings*, at the end of each of which the process begins all over again under identical conditions)?

In particular, since the alternation of strings in the two directions (i.e. in the positive and negative half-planes) is itself a Heads and Tails process when the strings are considered as 'tosses', it seems obvious that the balancing of periods in the lead should hold by analogy with the balancing up of the frequencies of Heads and Tails. In actual fact, this conclusion is true if one considers the *number* of strings giving the lead in one or other of the two directions, but it is *not* true if one wishes to consider the respective total *durations* of periods in the lead. In fact, we have seen (see the *Remark* in Section 8.7.6) that in an interval formed by complete strings (i.e. those ending in a zero) all durations are equally probable, instead of, as one might have expected, those of intermediate length being more probable (i.e. we have a distribution into almost equal periods). In the general case (where the final string may not be complete, i.e. the interval does not

end in a zero; see the *Remark* in Section 8.7.9), the situation is, in fact, the very opposite; it is the most unbalanced distributions which are most likely.

Both the form of the density, $f(x) = K/\sqrt{[x(1-x)]}$, and that of the distribution function, $F(x) = K\cos^{-1}(2x-1)$, show clearly that the extremely asymmetric values are favoured (although in a symmetric fashion so far as the two opposite directions are concerned). The best way to visualize the result is to note that the splitting of the total duration of a long game into the fractions x and $1-x$ of the duration in which one or other of the two gamblers is ahead can be thought of as brought about by choosing 'at random' (i.e. with uniform density) a point on the circumference of the semicircle having the segment [0, 1] as diameter, and then obtaining x by projecting that point onto this diameter. In other words, if the circumference is divided into an arbitrary number of equal arcs, their projections onto the diameter (which will clearly be smaller the nearer they are to the end-points) are equally probable (because they contain the point dividing the two parts x and $1-x$).

The reader should now examine Figures 8.9c, 8.9b and8.9a (in reverse order, from bottom to top), together with the notes which accompany it.

Feller (Vol. I, Chapter III) provides numerical data that illustrate this phenomenon and make clear why it is not, in fact, surprising. Imagine that two gamblers play continuously for a year (making a toss every hour, minute or second; it does not matter which). It turns out that there is only a 30% probability of both being ahead for more than 100 days (about 28% of the total time), whereas there is a 50% probability that one of them remains ahead for less than 54 days (15% of the time), 20% that he remains ahead for less than nine days (2·4% of the time), 10% for less than 2.25 days (i.e. less than 0·6% of the time – more than 99·4% for his opponent!).[36] Feller also provides the details of the behaviour resulting from a computer experiment.

8.8.2. It is not really surprising that these numbers are not what we would imagine intuitively. Intuition cannot guide us – not even roughly sometimes – in foreseeing the results from analyses of complicated situations. This is precisely why mathematics is so useful, particularly in probability theory.

We should ask ourselves, however, whether, even from a qualitative point of view, the above conclusions are paradoxical (and, if so, for what reasons), and how one might set about correcting and altering this impression by showing that it is, in fact, perfectly natural for things to be thus. Despite the fact that the example which has given rise to this discussion is an especially striking one, it is by no means a unique and isolated case and it provides us with an excellent basis for discussion and considerations relating more or less directly to more general problems. On the other hand, it is not so much the individual result itself that merits and requires illustration but rather the nature of random processes which – like the very simple case of Heads and Tails – are based on the simple idea of stochastic independence (or lack of memory, if one prefers to think of it in this way). Although this is a simple notion, it is difficult to understand it sufficiently well to avoid finding certain of its consequences paradoxical. We have already commented upon this on a number of occasions, some of which we recalled above (the laws

36 Sometimes people refer to 'the less fortunate' player. This is not quite right, however, since it is possible (although not very probable) that the one who has been in the lead for most of the time finds himself behind at the end (see Section 8.7.10).

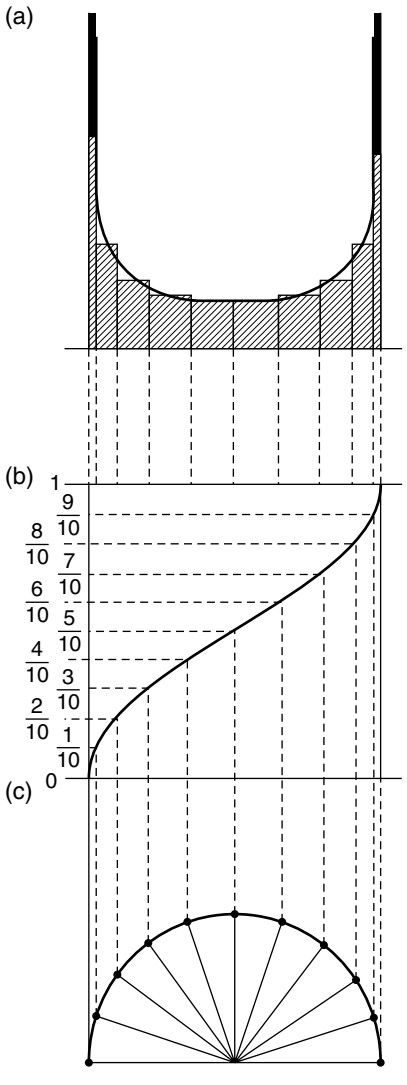


Figure 8.9 These should be read in reverse order (i.e. (c), then (b) and finally (a)).

a) The density of the arc sine distribution. The histogram shows the average density in each of the intervals between the deciles. The graph shows the density, whose equation, taking the interval to be [0, 1], is

$$f(x) = K / \sqrt{\left[x(1-x) \right]},$$

and is infinite at the end-points.

b) The distribution function of the arc sine distribution (obtainable using the device shown in (c)).

The abscissae shown correspond to the 'deciles' (see Chapter 6, 6.6.6) since they are obtained from the corresponding ordinates.

The ten intervals between the deciles are equally probable (with probability $\frac{1}{10}$). Note how much more dense the probability is near the end-points.

c) The probability distribution of the projection (onto the diameter) of a point 'chosen at random' (with uniform density) on the circumference.

This distribution occurs, for example, if one measures, at a 'random' instant, the position (or velocity) of a point performing harmonic oscillations.

The division of the circumference into 10 equal parts (18°) gives the deciles.

of large numbers and the long-expected time to ruin). These are – like the present example, and others which we shall soon come across – topics that are interrelated and deal with the same kinds of questions.

The reasons why these results appear paradoxical are all related to various kinds of distortion of the relations between probability and frequency:

by assuming connections without taking into account that they only exist under certain restrictive conditions;
by thinking that they virtually entitle one to make a prediction rather than a prevision;
by assuming that they systematically fall into familiar patterns of statistical 'regularity';
by having such a strong belief in such regularity as to make of it an autonomous principle; this leads one, inadvertently, to expect 'compensations' to take place in a

more extreme form and providential manner and sense than can be derived legitimately from probabilistic assumptions.

The danger of falling into these traps is even greater when one has been taught statistical concepts in a grossly oversimplified form, easily misunderstood and without the necessary warnings being given. The use of certain forms of terminology – for example, saying that something occurs *on average* a given number of times per unit time (instead of saying *in prevision*) – can lead one to regard such 'regularities' as certain, instead of merely being probable; that is, as predictions instead of previsions.

If 'regularity' is assumed as an 'article of faith' (and there is a book on statistics, inspired by this outlook, which is entitled 'Gleichförmigkeit der Welt'), how can it be, one might ask, that phenomena like returns to equilibrium and the distribution of leads could violate this regularity, thus challenging the supreme dictate of the ordered universe? If one thinks of returns to equilibrium (which are practically certain) as revealing a tendency towards, or desire for, such 'regularity', one would expect that a particle, having gone a long way below the origin while performing a random walk, would make an about turn in order to return to the fold. On the contrary, it has no memory and there is no fold to return to. It might carry on until it is twice as far from the origin before it turns back towards it, or it might have only gone half as far. If it does end up by going back to the origin with certainty, this is simply because, being on a random walk, it will sooner or later pass through all the points (but without any possibility of recognizing the point we have labelled the 'origin', nor any desire to do so).[37] One would have a stronger case (and, indeed, a valid one were it not for omitting to point out the fact that the expected duration is infinite, or, at least, for not taking it into account) if one were to argue that the phenomenon should reproduce itself 'regularly' because after each return to the origin a new string begins under precisely the same conditions.

If we attempt to identify and explain those reasons that we assume to underlie the tendency to talk in terms of 'paradoxes', we find the answers staring us in the face. The probability–frequency relation as it occurs in the law of large numbers should not be assumed to hold, since the successive Y_n are not independent. They depend upon a 'cumulative effect', which tends to be dominant; deviations take place only slowly and returns to equilibrium and changes of sign, that is of 'lead', only seldom. We have already mentioned above the idea of some kind of restoring force causing returns to equilibrium. However, only the last point is really important, because it pin-points a subtle and basic difference (whereas the other points simply caution one against the possibility of trivial and rather absurd misunderstandings).

The very fact that the probability of a return to the origin at time $t = n$ *tends to zero* (for n even, $u_n = 0{\cdot}8/\sqrt{n}$) should be sufficient to rule out the 'regularity' or 'stationarity' of behaviour that is the implicit and unconscious assumption occasioning all the 'astonishment' at these 'paradoxes'. The latter appear as such simply because they do not fit

37 I do not mean to imply that fallacious ideas of this kind are accepted statistical doctrine in some other approach differing from the one we follow. However, the environment created by a few introductory sentences followed by empirical clarifications etc. does not seem to be sufficiently antiseptic to prevent the germs of these dangerous distortions from multiplying in the subconscious.

into that particular framework; a framework that has created, in its own image and likeness, psychological attitudes whose tendency to rise to the surface becomes, in the absence of any process of re-education, general and indiscriminate. The following points may serve to provide a better appreciation of just how 'sensational' are the consequences of this probability tending to zero (a fact which is so simple when it is accepted as it is, without further thought). In a consecutive sequence of k tosses (e.g. 100 000), the probability that Heads always occurs from some n onwards is very small, but nonetheless finite; and this is also true for the probability that, starting from n, the sequence 11001… (where 1 = Heads and 0 = Tails) represents, in the binary code, either the first 100 000 decimal places of π, or the text of the Divine Comedy, or that it reproduces the initial segment obtained with the first k tosses, or any other preassigned segment of fixed length. Something of this kind occurs, in prevision, once every 2^k tosses, and it is practically certain that it occurs at least once within every segment of length N, if N is considerably larger than 2^k (and also that it occurs at least 10 times, or at least 1000 etc., provided we take a long enough sequence; the details are straightforward and we shall not bother with them here). On the other hand, the expected number of returns to equilibrium with a given number, N, of consecutive tosses starting from $t = n$ is, approximately,

$$\left(0 \cdot 8/\sqrt{n}\right).N/2 = 0 \cdot 4N/\sqrt{n} \to 0,$$

and the probability of at least one is even smaller. This means that, if one proceeds far enough to have an interval sufficiently long to give a non-negligible probability of containing a return to equilibrium (e.g. 1% or 10%), then one has an interval which almost certainly (e.g. with probability 90% or 99 %) contains the Divine Comedy at least once, and, if one carries on, at least 10 times, 1 million times, and so on, indefinitely.

8.8.3. Turning to a consideration of the values of Y_t (and not only at those instants for which $Y_t = 0$), a topic we shall be dealing with shortly, we can deduce an immediate and straightforward result about the extreme length of the strings at times far away from the starting time $t = 0$. We know that $|Y_n|$ has probability $\simeq 0{\cdot}8\ M/\sqrt{n}$ of being less than a preassigned M. For sufficiently large n, it is therefore almost certain that $|Y_n| > M$, in which case the string containing the instant $t = n$ *necessarily* has length $>2M$. In fact, the length would equal $2M$ under the assumption that the increase from the previous zero to Y_n and then the decrease to the following zero take place in an unbroken sequence of M successes and M failures, respectively. From considerations made in the context of the ruin problem, however, we know that it is probable that the increase and the decrease take much longer.

But the above remarks only have an illustrative and introductory value: they help us to see what is happening but they do not yet provide us with an explanation; neither do they resolve the confusion by going back to the source. At most, there is a restatement of the problem: instead of asking ourselves *why do the lengths of the strings become longer and longer* (despite the fact that they begin again from zero under the same conditions), we can ask why, given the same assumptions, do the ordinates Y_n become larger and larger (in absolute value); that is *why do the strings get further and further away from the axis* $y = 0$ (and it is clear that the two questions, even if not identical, are closely related).

Let us study then the history of each string. We might as well consider the first one, starting at $t = 0$. The probability that its length is $n = 2m$, that is that it terminates at the nth toss (with a return to equilibrium), is

$$u_n / (n-1) = u_{n-2} / n.$$

But u_{n-2} is the probability that no zero has previously occurred, so $1/n$ is the probability that the string terminates at time n *assuming that it does not terminate earlier* (for n even; otherwise, the probability is zero). It follows that a string has probability $\frac{1}{2}$ of terminating at the second toss, $\frac{1}{4}$ at the fourth (provided it did not terminate at the second), $\frac{1}{6}$ at the sixth (provided it did not terminate at the fourth) and so on. In demographic terms, a string can be thought of as an individual whose probability of dying decreases with age (this happens for newborn babies, the probability of survival increases each day they survive; the difference is that they grow old and the conditions become worse, whereas for the strings they continue to improve). We therefore see why the probability of 'the string enclosing the instant t' is smaller if t is smaller than if t is large. In the first case, it is necessarily 'young' (age at most t) and this bounds the past duration (the age that has been attained) and provides less favourable possibilities for the future (since an individual gets stronger with age): it has probability at least $1/(t + 2)$ of terminating at the next even toss, and so on.

And what are the probabilities of the various possible values for Y_n? They are no longer those of the Bernoulli distribution that we had before. We are now in a different *state of information*, because we are discussing a string; in other words, we know (or assume) that Y_t has not vanished in the meantime, and that the probability we seek is that conditional on this hypothesis, H, as we have already seen (Section 8.7.9(c)). This means that knowing that there are no zeroes modifies the distribution in favour of the larger values (as is natural); more precisely, it alters the probabilities in proportion to their sizes.

Every change in the state of information brings about a modification. For instance, if I knew the values of Y_n at every instant, then the probability of the string ending at the next toss – let us take n to be odd – is no longer $1/(n + 1)$, but is $\frac{1}{2}$ if $Y_n = \pm 1$, and zero otherwise. The situation would be different again if I knew just a *few* of the past values. If I knew Y_t at certain instants $t = n_1, \ldots, n_k$, the probabilities would be those conditional on the last value; that is on the hypothesis $H = (Y_{n_k} = c)$. But beware! This is only true if I have no knowledge, no clue whatsoever, relating to the results following the last known result. For example, if I obtain information every time $Y_t = 0$, I not only know the position of this last zero but I also know that no other zeroes have occurred since (and I then have the distribution given above, for the present Y_n, whereas, otherwise, I would have the Bernoulli distribution). Yet another situation would arise if I knew it to be more probable for information to be available in the case of returns to the origin, or in the case of large values being attained, and so on, than in other cases: *absence of information can itself be informative*. In the cases just mentioned, it increases the probability of the nonoccurrence of those things which, had they occurred, would probably have been reported.

Many of the mistakes that are made in the probabilistic treatment of problems and phenomena derive precisely from either ignoring, or forgetting, or giving insufficient weight to, the following fact: that everything depends upon the current, actual state of

information (with all the attendant flexibility that attaches to this notion in practice). In interpreting deterministic laws, also, we need to keep circumstances of this kind in mind. An example is provided by the treatment of 'hereditary' phenomena, such as hysteresis, using integral or integral–differential equations. If one assumes that knowledge of the past enters the picture indirectly through the modification it produces in the structure determining the present state (which is itself not directly observable), then we see that certain information (in our case, 'the past') may or may not be 'informative' in so far as the effects we are interested in are concerned (here, *deterministic* prevision of the future – i.e. 'prediction' – rather than prevision), depending on whether certain other information is, or is not, available (here, complete information about the structure of the 'present state'). If this latter information is not available, the information concerning the past serves as a substitute. It may be a completely adequate substitute, or only partially so, according to whether the knowledge of the outside influences in the past are, or are not, considered sufficient to determine, completely and with certainty, the present unobservable situation. In the latter case, we are essentially back in the realm of probability, even if this remains obscured by the fact that the treatment deals only with macroscopic behaviour, neglecting the random aspects which are, in that context, negligible.

In the probabilistic field, however, information is always incomplete and derives from the distinction – which, in any case, is never very clear-cut – between what one knows, or believes one knows, or definitely remembers, and what one does not know. A fundamental rôle is played by that certain something which is, in a sense, complementary to information, and which comes about by interpreting the reasons for the absence of information. We have illustrated this with the Heads and Tails process, and we shall return to it again and again, sometimes with illustrations which are particularly instructive because they are, at first sight, rather disconcerting.[38] For examples from more familiar fields, note that the knowledge of a person's age is, to some extent, a substitute for a medical examination in so far as the evaluation of the probability of death is concerned and that, for a person who is insured, present age, together with the medical report dating back to when the policy was originally taken out, are taken as substitutes for a medical examination at the present time. In like manner, the fact of whether or not one receives direct news, or whether or not the newspapers carry reports of a certain situation, individual, firm, institution and so on, itself constitutes information (satisfactory or not, as the case may be). Any attempt – and these are still frequent – to base the theory of probability on some distinction between those things of which one is *perfectly certain* and others of which one is *perfectly ignorant* precludes, for the reasons we have given (and by not taking into account objections of principle), an understanding of the most meaningful aspects of problems requiring the use of probability theory.

Going back to the discussion of the questions we considered for the Heads and Tails process and to the doubts expressed in this context ('*How can it be…* ?'), we see, therefore, that, in line with what has been said, the answer lies in making it clear that the situation – for example, the probability of a return to equilibrium at a given instant – does not alter merely because of the passage of time, or because time modifies something or other, but rather because our state of information changes. Initially, that is not

38 For example, the equivalence of the Bayes–Laplace scheme to that of Pólya's 'contagion probabilities'; see Chapter 11, 11.4.4.

conditional on any subsequent information, our state of information consists solely of the knowledge that $Y_0 = 0$. When we later place ourselves at times $t = n$, however, we have a *changed* state of information; one that consists in knowing that there was a passage through the origin *n steps ago* (without knowing whether or not this was the last zero, nor anything else which would lead us to depart from an identical state of information regarding the 2^n possible paths that could have been followed meanwhile).

8.8.4. Despite all this, the doubt might linger on, transplanted to the new ground opened up by the new information. How can it be that the variation in the state of information to which we referred continues to exert an influence even when we move indefinitely far away? There is, in fact, a case in which knowledge of the initial state, provided it is sufficiently remote, ceases to have any influence (this is the so-called *ergodic* case, which we shall be concerned with briefly in Chapter 9, 9.1, when we deal with 'Markov Chains'). This is, however, something that only occurs under certain specified conditions and the fact that it often crops up is merely an indication that these conditions are satisfied in many of the problems one considers, rather than because of some principle which permits one to use the idea indiscriminately.

Here, too, as in the case of belief in a tendency to equilibrium, it happens that a special circumstance is assumed as some kind of autonomous 'principle', rather than as a simple and direct consequence of conditions that may hold in some cases and not in others. In this way – partly by accident and partly as a result of the usual obsession with replacing probability theory by something which is apparently similar, but which can, in fact, be reduced to the ordinary logic of certainty – one ends up by seizing upon the most fascinating results (like the laws of large numbers and the ergodic theorem) and raising them to the status of principles. When the applications of these principles to situations in which the theorems they misrepresent do not hold turn out to be contradictory, the results are then held to be paradoxical. As an analogy, it is as if, instead of the principle of conservation of energy, one took as a 'principle' the statement that a field of forces must be conservative, and then were faced with justifying the 'paradoxical' cases (like magnetic fields) where the 'principle' no longer holds.

As an example, it is often asserted – especially by philosophers – that the calculus of probability *proves* that the 'ergodic death' of the universe is inevitable. On the contrary, the calculus of probability (the logic of uncertainty) is completely neutral with respect to facts and behaviour relating to natural phenomena, and with respect to any other kind of 'reality' (in just the same way as the logic of certainty is). It is absurd to believe that the calculus of probability can itself rule out any particular belief or that it can force one to adopt it; whether it be a belief in 'ergodic death', or whatever. All that it does do is to rule out 'incoherent previsions', on the grounds that these are not previsions; in the same way as the logic of certainty precludes one making the assertion that a horse has three fore legs and four hind legs, making a total of five (whereas it would be admissible to say $3 + 4 = 7$, or $3 + 2 = 5$, or $1 + 4 = 5$). The point is that it must not be inconsistent; the question of whether or not the statements conform to what zoologists regard as admissible is not relevant. Ergodic death is very probable if one accepts, or at least assumes as the most plausible, that model of physical phenomena which regards things as deriving from the destruction of an initial state of order (as in the mixing of gases; kinetic theory). But the calculus of probability in no way precludes phenomena in which a new order is created (as in biology; in particular, the mechanisms of reproduction for

DNA, and hence of cells, human beings, and – who knows? – new stars or galaxies);[39] on the contrary, its techniques provide the means of analysing them.

Returning to our case, one could say that the 'ergodic principle' *no longer applies* (which is another way of saying that the 'ergodic theorem does not apply unless the necessary assumptions are satisfied'), because if, at time $t = n$, we know that $Y_0 = 0$, then, amongst other things, we already know that $|Y_n| \leqslant n$ with certainty (not to mention our knowledge of the distribution); this information is significant, although its significance varies a great deal with n. The opposite situation – the ergodic case – occurs if we think of the same random walk on an m-sided polygon (m odd; a step clockwise or anticlockwise, according to whether we get a Head or a Tail). It is clear that knowledge of the starting point is practically irrelevant for evaluating the probabilities of the m positions after n steps (for n large); these probabilities will all be practically equal (to $1/m$).

There is one thing that we have seen, however, which seems to contradict the (obvious) fact that the process begins again under identical initial conditions after every return to zero. If this is so, how is it that the first zeroes can be expected to be very close and then subsequently get further and further apart in the startling way brought home by our discussion of the interposed repetitions of the Divine Comedy? There is, in fact, no contradiction. Each time a zero occurs, it is to be expected that there will be several close to each other; the initial period is a special case of this. As a result, obviously, the groups of zeroes are even further from one another than the individual zeroes would be if we took the same number of them but assumed them approximately equidistant. For an arbitrary zero, for example, the kth, conveying *no information* about the length of adjacent strings (as would be the case if one said, for instance, 'the first zero after the nth toss', because it is likely that the n*th* toss will fall within a long string), the probability is always $\frac{1}{4}$ that the two adjacent zeroes are the minimal distance away (=2; in other words, that the two adjacent strings have the minimal lengths possible, i.e. 2); $\left(\frac{5}{8}\right)^2 \simeq 0 \cdot 39$ that they are both not more than a distance of 4 away; $\left(\frac{193}{256}\right)^2 \simeq 0 \cdot 75$ for not more than 10 away; and so on (in general, the probability is $\left(1-u_n\right)^2 \simeq 1-2u_n \simeq 1-1 \cdot 60/\sqrt{n}$ that both adjacent strings have lengths not exceeding n). Briefly, there are a great number of short strings, but here and there we find long strings, some extremely long; when we count not the numbers of strings but the number of steps they contain, however, the proportionate contribution from short and long strings is inverted. This is what we saw (see the *Remark* in Section 8.7.6) when we compared the probabilities u_{n-2}/n with the previsions of the lengths $n(u_{n-2}/n) = u_{n-2}$. In the context of this conceptual discussion, it is convenient just to take up the final point (mentioned above) concerning the trouble caused by the expression 'on average', a notion inspired by the statistical formulation.

Once again, we are dealing with the attempt to replace a genuine and valid probabilistic concept, which applies under all circumstances, by a counterfeit notion, only partially valid and not always applicable (it does not apply here, for example).

---

39 For a brief summary of how 'chance' comes to intervene continuously in thousands of complicated ways to bring about evolution (albeit, of course, according to our present conceptions), it is sufficient to read the two sections entitled 'The Development of Life' and 'A Chance Happening' in V.F. Weisskopf, *Knowledge and Wonder*, Heinemann, London (1962). As for the 'continuous creation of matter' and the formation of the galaxies, see the section entitled 'What happened at the beginning?', pp. 165–167; also see D.W. Sciama, *The Unity of the Universe*, Faber and Faber, London (1959); in particular, the section on the 'Steady State Model', pp. 155–157.

The probabilistic meaning is expressed perfectly – even though the expression may be rather unpalatable – by saying that each *individual* string has an infinite *expected length*, which is a result of the possible lengths 2, 4, 6, 8,... having probabilities $\frac{1}{2}, \frac{3}{8}, \frac{5}{16}, \frac{35}{128}$, ..., respectively. It makes no difference (just as it makes no difference whether we quote an interest rate as 45 lire per 1000 lire, or as 0–045) if we say that in 1000 strings we have, in prevision, a total length of 1000 deriving from strings of length *two*, 375 from strings of length *four*, 312·50 from length *six*, 273·44 from length *eight* and so on. There is no harm in this, and, indeed, it could be more expressive to consider that an expected length of 312·50 steps from 1000 strings derives from an expected number of strings of length six equal to 52·087. The trouble arises when one tries to interpret the phrase in a nonprobabilistic sense, as if it were possible to state that in any 1000 strings *things will turn out in this way* (in some vague sense: no-one goes so far as to claim this to be logically true – i.e. in a definite necessary sense – but people omit to state that it is at most 'very likely'; it is as though the possibility of something else happening could be avoided by recourse to some hybrid notion of 'practically certain'). Anyway, this very fundamental objection, one which always applies, precludes one from making statements of this kind without qualifying them as being *almost certain*, where by *almost-certainty* – by the mere fact of it not being *certainty* – we mean simply a rather high degree of probability (the latter being always subjective).

But it is not enough merely to correct a conceptually and formally inadequate form of expression. We must also make clear that statements which assert that in a large number of trials (in our case, strings) the actual outcomes will very likely be close to the 'previsions' do not hold except under appropriate conditions. First of all, we need conditions like independence, and this holds in our case; so far as the lengths and the signs are concerned, the strings are stochastically independent. For this reason, we can conclude that we may be almost certain that the proportion of positive and negative strings is fifty-fifty; the sequence of strings thought of in terms of their signs is a Heads and Tails process. However, we cannot claim that the same is true for the number of steps made on the two half-planes: despite the independence, the conclusion fails to hold because the prevision of the length of a string is infinite. *A fortiori*, for precisely the same reason, we cannot make statements of almost certainty about the frequency distribution of the lengths of the strings (or of the number of steps, considered in terms of the length of the string to which they belong). On the contrary, it would be very difficult to formulate this problem (even if the expected length were finite, and the statement therefore essentially true), not least because there are an infinite number of cases (lengths) to be distinguished.

We have shown how even a fairly superficial examination of very simple cases, like that of the Heads and Tails process, can reveal a number of features which are both unsuspected and fascinating in their own right. The intrinsic interest of these results has interesting conceptual implications when one considers more deeply the reasons for the surprise and the air of paradox which they generate.

8.9 Properties of the Wiener–Lévy Process

8.9.1. In Section 8.3, we had a brief look at those properties of the Wiener–Lévy process that could be established immediately and which served to enable us to make reference to the process. We now return to this topic, both in order to consider it in more depth and to show how certain asymptotic properties, which hold in many cases of

asymptotically normal processes, gain in simplicity and clarity when we observe that they are exact properties in the Wiener–Lévy case.

This is, paradoxically, both the simplest and the most pathological case. The disarmingly simple features we have seen already: all the quantities we consider, whether individually, in pairs, or n at a time, and under all the circumstances we have examined, are normally distributed (in either 1, 2 or n dimensions). What could be better?

There was one feature, however, which might perhaps have given us grounds for suspecting that troubles might lie ahead. We are referring to the property of projective invariance, which, as we mentioned, enables us to reduce the study of asymptotic behaviour to that of the behaviour in the neighbourhood of the origin. At the time, we did not wish to frighten the reader by drawing attention to certain things that happen at infinity and cause awful trouble when one considers them concentrated near the origin. What is even worse is that everything that happens in the neighbourhood of the origin also happens in the neighbourhood of every other point of the curve $y = Y(t)$ (since the process is homogeneous with independent increments).

8.9.2. First of all, we recall how the Heads and Tails process (along with many others) provides – with an appropriate change of scale – a good approximation to the Wiener–Lévy process (to any desired degree of accuracy).

Let us consider the standardized Wiener–Lévy process ($m = 0$, $\sigma = 1$). A Heads and Tails process, in order to preserve these characteristics and to approximate the continuous process, must consist of a large number of small jumps (e.g. very frequent bets with very small stakes) and, instead of a single jump +1 per unit time, requires N^2 jumps of size $\pm 1/N$ per unit time. In this way, the standard deviation per unit time is, in fact, given by

$$\sigma\sqrt{n} = (1/N).\sqrt{N^2} = 1$$

as required.

If N is large – in the sense that the time intervals $\tau = 1/N^2$ and stakes $s = 1/N$ are small in comparison with the precision with which we wish, or are able, to measure intervals of time and amounts of money – this process is practically indistinguishable from the Wiener–Lévy process. In fact, if all the time intervals we wish, or are able, to consider contain a large number of small time intervals, τ, then the increments of $Y(t)$ are made up of the sums of a large number of independent increments and are, therefore, approximately normally distributed.

If we think in graphical terms, we can say that if the graph of the Heads and Tails process (Figure 7.1 in Chapter 7, 7.3.2) is collapsed by dividing the ordinates by $1/N$ and the abscissae by $1/N^2$, with N large enough to render imperceptible the segments of the broken line corresponding to the individual tosses, then we have the most exact obtainable representation of a Wiener–Lévy process. In a certain sense, this is precisely the same old process of approximation and idealization as is used when we consider changes in population (the number of inhabitants of some region etc.) as a continuous graph: even though one wishes to consider them as drawn in, one chooses the scale in such a way as to render imperceptible the jumps that represent the individual births and deaths on which the behaviour of the curve actually depends.

8.9.3. Of course, as we remarked at the time, instead of starting from the Heads and Tails process in discrete time, we could start from that in continuous time (the Poisson

variant, with jumps $\pm 1/N$, N^2 of them, *in prevision*, per unit time), or with any other distribution of jumps (e.g. normal, always taking the standard deviation to be $1/N$ etc.).

Conversely, the Wiener–Lévy process can be a useful representation (giving an excellent approximation on some given scale) of phenomena whose 'microscopic' behaviour may well be very different. Among other things, it provides a useful model of the *Brownian motion* of a particle (or, better, if we restrict ourselves to one dimension, the projection of its motion onto one of the axes). Of course, the scale must be chosen so that it no longer makes sense to attempt to follow the actual mechanism of the phenomena, with its free paths, collisions and so on. We also note that P. Lévy often refers to the Wiener–Lévy process as the 'Brownian motion process' (the name Wiener–Lévy has come about in recognition of the two authors who made the greatest contributions to the study of this process; Bachelier also deserves a mention, however – he had previously discovered many of the properties and results, although in not such a rigorous manner).

8.9.4. We shall restrict ourselves in what follows to collecting together, as a survey, some of the more interesting facts about the process, but without providing proofs. In general, however, we shall be dealing with results that have already been proved implicitly – or at least made plausible – by virtue of results established for the Heads and Tails process.

Problems relating to the Wiener–Lévy process can be tackled in many different ways; in a certain sense, this reflects the various ways of looking at the normal distribution, which we noted when the latter was first introduced (the Wiener–Lévy process can be considered, roughly speaking, as a particular form of extension of the normal distribution to the infinite-dimensional case). On the other hand, we get a better overall view if we say something briefly about each of the most important procedures.

Those procedures, which derive basically from the Heads and Tails process, or something similar, are essentially rooted in combinatorial theory (together with whatever else may be required). The greater part of Chapters 7 and 8 are, in fact, devoted to this kind of procedure and we have often pointed out how it might be used in the context of the Wiener–Lévy process. We shall shortly give the details of this.

Rather more direct procedures are derived from the properties of the normal distribution itself, together with the various techniques for dealing with distributions. This means that knowledge of the second-order characteristics (variances and covariances) are sufficient to determine the process. We have already given examples of this when introducing the preliminary properties of the Wiener–Lévy process.

The third kind of procedure will require a more thorough discussion. It involves the study of diffusion problems (the heat equation and so on) and was briefly mentioned in Chapter 7, 7.6.5, and again in Chapter 8, 8.6.7. Despite the elegance and power of these methods, we shall not be able to say a great deal about them here. This is unfortunate, because, besides their power, they can be made very expressive in terms of the image of the spread of probability, considered as mass. However, we clearly cannot include everything and this seemed a reasonable candidate for exclusion, as – from a conceptual viewpoint – it is more in the nature of an analogy than a genuine representation of the problems. This is in contrast to the other approaches, which, in their various ways, stick closely to the probabilistic meaning and enable one to shed light on every aspect of it.

Anyway, we shall restrict ourselves here to illustrating, using the perspective provided by diffusion theory, some of the problems with which we are already familiar through other approaches.

The Wiener–Lévy process corresponds precisely to the basic case of diffusion starting from a single source. The so-called 'dynamic' considerations and conclusions, in which t is considered as the time variable (instead of just a constant), involve precisely this process (and not just the individual distributions at individual instants).

The gambler's ruin problem (in the version provided by the Wiener–Lévy process) requires the introduction of an absorbing barrier; the straight line $y = -c$, where c = the initial capital. This problem can be solved, in the theory of heat transfer, by the *method of images* (due to Lord Kelvin). This involves placing an opposite (cold) source at the point $t = 0$, $y = -2c$ (a mirror-like image of the origin, a hot source, the image being taken with respect to the barrier). The resulting process, which, for reasons of symmetry, clearly has zero density on the barrier, gives, at each instant t, the density of the distribution of the gain. The missing part (the integral of the density is less than 1) is the mass absorbed by the barrier; that is the probability of ruin before the instant under consideration. It can be seen, without the need for any calculations, that this is twice the 'tail' that would go beyond the barrier (this tail is itself missing and there is also the negative tail that has come in from the cold source). We note that this corresponds precisely to Desiré André's argument.

Similarly, in the case of the two-sided problem, the method of images leads to the introduction of an infinite number of hot and cold sources (images of the actual source, with an even or odd number of reflections in the absorbing barriers). This is the 'physical' interpretation of the formulae in Section 8.6.6.

8.9.5. Our survey of the results relating to the Wiener–Lévy process should begin, naturally enough, with those we gave in Section 8.3, and with those we came across subsequently. We shall only repeat those things which are required to make the survey sufficiently complete.

We begin with the results relating to the ruin problem (i.e. to the case with an absorbing barrier).

In the case of a single barrier (at $y = c$, say), the probability of ruin at or before time t, $F(c, t)$,[40] that is the distribution function for the time T spent by the process before ruin occurs, $F(c, t) = P(T \leqslant t)$, is given by

$$\begin{aligned} F(c,t) &= \mathbf{P}\left(\left|Y(t)\right| > |c|\right) = 2\mathbf{P}\left(Y(t) > |c|\right) \\ &= \frac{2}{\sqrt{(2\pi t)}}\int_{|c|}^{\infty} \mathrm{e}^{-y^2/2t}\,\mathrm{d}y = \frac{2}{\sqrt{(2\pi)}}\int_{|c|/t}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \end{aligned} \tag{8.47}$$

(the above holds in a real sense, because the probability of $T < \infty$ is 1). The density has the form

$$\frac{\partial F}{\partial t} = f_c(t) = \frac{|c|}{\sqrt{(2\pi)}} t^{-\frac{3}{2}}\ \mathrm{e}^{-c^2/2t} = \frac{|c|}{Nt}\cdot\frac{N}{\sqrt{(2\pi t)}}\mathrm{e}^{-c^2/2t}. \tag{8.49}$$

40 In Sections 8.6.4–8.6.5, this was denoted by $q(t)$ and $p(c)$ (or $q_c(t)$ and $p_t(c)$) because it was then convenient to think of one of the variables as fixed (i.e. included as a parameter).

We recall that we are dealing with the stable distribution with index $\alpha = \frac{1}{2}$. The second form emphasizes the relationship with the Heads and Tails process, involving N^2 tosses per unit time, each involving a gain of $\pm 1/N$. The second factor expresses (asymptotically) the probability of a gain of $|c| = (N|c|)(1/N)$ in N^2t tosses, the first factor the probability that we are dealing with the first passage through level $y = |c|$ (see Section 8.7.9(*c*)).

When considered as a function of $|c|$ (and we shall write y rather than $|c|$), the distribution becomes the half-normal, with density

$$f_t(y) = \sqrt{(2/\pi)}\,e^{-y^2/2t}\ \left(y \geqslant 0\right)\left(\text{i.e. zero for } y < 0\right). \tag{8.48}$$

We recall that, in addition, this holds for the following cases:

the absolute value of $Y(t)$
the absolute value of $\vee Y(t)$ (the maximum of $Y(\tau)$ in $0 \leqslant \tau \leqslant t$)
the absolute value of $\wedge Y(t)$ (the minimum of $Y(\tau)$ in $0 \leqslant \tau \leqslant t$)
the absolute value of $\vee Y(t) - Y(t)$
(the deviation from the maximum)
the absolute value of $\wedge Y(t) - Y(t)$
(the deviation from the minimum).

We now give the probability distributions of $Y(t)$ conditional on three different assumptions concerning the maximum of $Y(\tau)$ in $[0, t]$. The three assumptions are as follows:

that, with respect to some given $c > 0$, we have $\vee Y(t) \geqslant c$ (in equation 8.75);
that $\vee Y(t) \geqslant c$ (in equation 8.76);
that $\vee Y(t) = Y(t)$(in equation 8.77).

In the first two cases we have:

$$f(y) = K \exp\left\{-\left(c + |y - c|\right)^2 / 2t\right\}, \quad 1/K = F(c,t), \tag{8.75}$$

$$f(y) = K\left[\exp\left\{-y^2/2t\right\} - \exp\left\{-(2c - y)^2/2t\right\}\right](y \leqslant c), \quad 1/K = 1 - F(c,t). \tag{8.76}$$

The first follows immediately from the reflection principle. Note that $c + |y - c|$ is equal to y (for $y \geqslant c$) or $2c - y$ (for $y \leqslant c$), and that the distribution is therefore the normal with the central portion (between $\pm c$) removed, and the two remaining tails attached to one another. For the second, it is sufficient to observe that, multiplying it and the first one by $1 - F(c, t)$ and $F(c, t)$, respectively (i.e. by suppressing the K), and summing them, we must again obtain $K \exp(-y^2/2t)$.

Finally, suppose we assume that we know that either the value $Y(t)$ is greater than all those previously obtained, that is that $Y(t) = \vee Y(t)$ (without knowing anything more about the actual value), or that we know that $\wedge Y(t) = 0$, that is that the minimum is the initial zero, $Y(0) = 0$ (by the reversal principle, the two cases are equivalent).

Conditional on either of these, the density of the distribution is like $\partial F/\partial t$ in equation 8.49, except that we now have to take $y = |c|$ as the variable rather than t. This changes K and we obtain

$$f(y) = \frac{y}{t} e^{-y^2/2t} \left(y \geqslant 0 \right), \tag{8.77}$$

giving the distribution function

$$F(y) = \left[1 - e^{-y^2/2t} \right] \left(y \geqslant 0 \right). \tag{8.78}$$

The same thing holds, with the range of values reversed, when we take $Y(t)$ to be equal to the minimum rather than the maximum (or if $Y(0) = 0$ is the maximum).

This can be justified by considering (in a somewhat roundabout manner) the meaning of $\partial F/\partial t$, or, alternatively, by considering equation 8.71 of Section 8.7.9, which refers to the Heads and Tails process.

In the case of two barriers (the ruin problem for two gamblers), the distribution of $Y(t)$ conditional on the fact that neither has been ruined in $[0, t]$, that is conditional on

$$-c' \leqslant \wedge Y(t) \leqslant \vee Y(t) \leqslant c'' \qquad \left(\text{with } c' > 0 \text{ and } c'' > 0 \right),$$

is given by

$$f(y) = K \sum_{h=-\infty}^{+\infty} \left[\exp\left\{ -\left(y - 2hc^* \right)^2 / 2t \right\} - \exp\left\{ -\left(y + 2c' + 2hc^* \right)^2 / 2t \right\} \right], \tag{8.79}$$

where $c^* = c' + c''$, and $-c' \leqslant y \leqslant c''$.

In the symmetric case, $c' = c'' = c$, $c^* = 2c$, this becomes

$$f(y) = K \sum_{h=-\infty}^{+\infty} (-1)^h \exp\left\{ -\left(y - 2hc \right)^2 / 2t \right\} \qquad \left(-c \leqslant y \leqslant c \right). \tag{8.80}$$

Clearly, the probability, $1 - q(t)$, of neither barrier being crossed until time t is equal to $1/K$ (which is given by the integral of the Σ between $\pm c$, or, in the general case, between $-c'$ and $+c''$). This $q(t)$ also appeared, in a slightly different form, in equation 8.53 (see Section 8.6.5). Note that, if we ignored the K, equation 8.79 would give $f(y)\,\mathrm{d}y$ = the probability that $Y(t)$ lies in $[y, y + \mathrm{d}y]$ *and* has never previously gone outside the interval $[-c', c'']$: the point is that we would be saying 'and', rather than 'assuming that'. Similar comments apply in all other cases of this kind.

The few remarks we have made about Lord Kelvin's 'method of images' (Section 8.6.7) suffice to explain the result. If one so wished, one could verify it by checking that both the diffusion equation (equation 8.32 of Chapter 7, 7.6.5) and the boundary conditions, $f(y) = 0$ on the half-lines $y = -c'$ and $y = c''$ for $0 < t < \infty$, are satisfied.

8.9.6. In the case of the Wiener–Lévy process, we can provide complete answers to questions concerning the asymptotic behaviour of $Y(t)$ as $t \to \infty$. In principle, these answers are provided by a celebrated result of Petrowsky and Kolmogorov; 'in practice', that is in a less complete but more expressive way, they are given by a famous theorem of Khintchin, the so-called 'law of the iterated logarithm' (see the brief comment in Chapter 7, 7.5.4).

What we do is to compare $Y(t)$ with some function $\omega(t)$ (which we assume to be continuous, increasing and tending to $+\infty$), and then calculate the probability that the inequality $Y(t) < \omega(t)$ holds from some arbitrary time t onwards. More precisely, we examine the limit, as $t' \to \infty$, of the probability that the inequality holds from t' onwards. To be even more precise,[41] this latter probability is itself to be understood as the limit, as $t' \to \infty$, of the probability that the inequality holds in (t', t''). The p of interest is thus given by

$$p = \lim_{t' \to \infty} \left[\lim_{t'' \to \infty} p\left(t', t''\right) \right];$$

the limit certainly exists, since $p(t', t'')$ increases as t'' increases and decreases as t' increases.

We can say a great deal more, however. By the Zero–One law, only the two values $p = 0$ or $p = 1$ are possible: either it is practically certain that $Y(t)$ remains below $\omega(t)$ from a certain time onwards, or it is practically certain that this does not happen; that is there will always be segments in which $Y(t)$ is greater than $\omega(t)$. The class of functions $\omega(t)$ can therefore be divided into two subclasses, which could be said to contain 'those which increase more (less) rapidly that the "large values" of $Y(t)$' The general distinction (given by Petrowsky and Kolmogorov) is that $\omega(t)$ belongs to the upper or lower class according to whether the improper integral (from an arbitrary positive t_0 to $+\infty$) of

$$\psi(t).t^{-1} \exp\left\{-\frac{1}{2}\psi^2(t)\right\} \mathrm{d}t, \quad \text{where } \psi(t) = \omega(t)/\sqrt{t}, \tag{8.81}$$

converges or diverges.

In terms of $\psi(t)$, the condition $Y(t) < \omega(t)$ can be written as $Y(t)/\sqrt{t} < \psi(t)$; that is in terms of a *standardized* function (for which we have made σ = constant = 1.

The more expressive distinction (that of Khintchin) simply considers the class of functions

$$\omega(t) = k\sqrt{(2t \ \log \ \log t)} \qquad \text{(i.e.} \quad \psi(t) = k\sqrt{(2 \log \log t)}) \tag{8.82}$$

41 This further qualification is unnecessary if countable additivity is assumed. We recall similar caveats in the case of the strong law of large numbers (Chapter 7, 7.7.3), etc. For simplicity, we shall give an informal discussion here.

and asserts that these belong to the lower class when $k \leqslant 1$, and to the upper class when $k > 1$. The result can be strengthened by considering the functions

$$\omega(t) = \sqrt{\left[2t\left(\log\log t + k\log\log\log t\right)\right]}: \tag{8.83}$$

these belong to the lower class if $k \leqslant \frac{3}{2}$, and to the upper if $k > \frac{3}{2}$ (a generalization due to P. Lévy, which is proved by first using a direct approach, which leaves the cases $\frac{1}{2} \leqslant k \leqslant \frac{3}{2}$ undecided, and then removing the gap by using the Petrowsky and Kolmogorov result).

The proof of the general criterion is based on diffusion theory ideas (which relate to the theory of heat flow). For the law of the iterated logarithm, we recall the previous comments given for the case of Heads and Tails.

8.9.7. *Small-scale behaviour* is extraordinarily complicated and irregular. Not only do all the large-scale peculiarities reappear (shrunk by a factor of N^2 in the abscissa, corresponding to a factor of N in the ordinate) but, also, if we study the behaviour in the neighbourhood of a point – the origin, for instance – we find all the asymptotic properties corresponding to $t \to \infty$ reappearing in an inverted way. This can be seen most simply by observing that, if $Y(t)$ is given by a Wiener–Lévy process, then the same is true for the function

$$Z(t) = tY(1/t);$$

this has $m_t = 0$, $\sigma_t = t\sqrt{(1/t)} = \sqrt{t}$, the distribution is normal, and the correlation coefficient between $Z(t_1)$ and $Z(t_2)$ is the same as that between $Y(1/t_1)$ and $Y(1/t_2)$ (if $t_2 > t_1$, and hence $1/t_1 > 1/t_2$, it is equal to $\sqrt{[(1/t_1)/(1/t_2)]} = \sqrt{(t_2/t_1)}$); this is all we need.

It is practically certain, therefore, that in every neighbourhood of zero ($Y(0) = 0$) $Y(t)$ vanishes *an infinite number of times* (as in the case when $t \to \infty$) and that it touches infinitely often every curve

$$y = \omega(t) = k\sqrt{\left[2t\log\log(1/t)\right]}$$

with $k \leqslant 1$, but not those with $k > 1$ (which gives, *locally*, an almost certain 'modulus of continuity'; $|Y(t_0 + t) - Y(t_0)| < \omega(t)$ in a neighbourhood of t_0, with $0 < t < \varepsilon$). If, however, we want this to hold almost certainly for all the t_0 of some given interval simultaneously (still for all t between 0 and ε), we have to take

$$\omega(t) = k\sqrt{\left[2t\log(1/t)\right]}, \qquad k > 1 \tag{8.84}$$

(the simple rather than the iterated logarithm).

In order to be brief, the presentation of the results in these cases has been rather informal. We should point out, however – for reasons we shall see shortl y– that there are grave dangers in treating these topics without sufficient care and attention. For every point t_0 at which $Y(t_0) = 0$, it is practically certain (probability = 1) that there are other roots (an infinite number of them) in every interval of the point, either to the left or to the right (and the same holds true at every other point if we consider crossings of the horizontal line $y = Y(t_0)$). On the other hand, between two roots there are always

several intervals (and almost certainly a countable infinity) in which $Y(t)$ is either positive or negative; hence there are isolated roots to the right or to the left – the end-points of such intervals. Since this can be repeated for all horizontal lines y = constant (an *uncountably infinite* set), the points $y = Y(t)$, which are isolated (on at least one side) from points of the curve at precisely the same level y, form, in every interval, an uncountably infinite set and among them there are always an infinite number of points isolated on either side (at least the maxima and minima).

This having been said, the length of the segment starting from the origin, where we assume $Y(0) = 0$ (or starting from some arbitrary t' at which we know that there is a root, $Y(t') = 0$), and containing no roots, is a random quantity X, which has probability 1 of being precisely zero (if 0, or, in general, t' is a root which is adherent on the left to the set of roots). Indeed, such a random quantity, $X(t')$, can be considered, without changing the problem, for any arbitrary t' – even if $y' = Y(t')$ is not zero – as the length of the interval on the left of t' not containing points t at which $Y(t)$ again takes the value y'. In any case, we know that we necessarily have $X(t') > 0$ for an uncountable infinity of points in any arbitrarily small interval, and it can be shown that, *assuming* the length X to be greater than some given $x_0 > 0$, the probability of it being greater than some $x \geqslant x_0$ is $\sqrt{(x/x_0)}$. In other words, conditional on the hypothesis $X \geqslant x_0$ ($x_0 > 0$), we can say that X has distribution function and density given by

$$F(x) = 1 - K\sqrt{x}, \tag{8.85}$$

$$f(x) = \tfrac{1}{2}K/\sqrt{x}, \tag{8.86}$$

where $K = 1/\sqrt{x_0}$ (so that $F(x_0) = 0$); as $x_0 \to 0$, we also have $K \to 0$.[42]

The result for Heads and Tails (where $X \geqslant n$ has probability $u_n \simeq 0{\cdot}8/\sqrt{n}$[43]) corresponds to the case $K \neq 0$ (because, clearly, in discrete time there is no way for 'peculiarities on the small scale' to occur).

---

42 This means not only that, in the absence of any contrary hypothesis (the case of an infinite number of roots adherent on the left), there is a probability = 1 that $X = 0$ (i.e. X is concentrated at the point $x = 0$), but also that with the single assumption that $x > 0$, all the probability is *adherent* to zero (i.e. however $x_0 > 0$ is chosen, the probability that $X \geqslant x_0$ is zero; this is obvious, because, in any finite interval, however large, there can only be a finite number of intervals containing no roots and of length greater than x_0, whereas there are an infinite number of 'small' intervals containing no roots in every interval of almost all the roots; i.e. excluding the isolated ones).

Note, of course, that the problem would be different if we were talking about an interval containing no roots, and chosen by picking out some point in it. As usual – recall 'sums' at Heads and Tails, 'number' and 'length' of strings, etc. – this procedure would favour the choice of the longest intervals (see the comments to follow in the text). The choice must be made by saying, for example, 't' = the starting point of the third interval of length $\geqslant x_0$ (possibly with some additional complications, in order that the restriction to a simple example is seen not to be necessary) after the level $Y(t) = c$ has been reached'. For the case $x_0 \to 0$ (under the assumption $X > 0$), this explicit method of choice does not exist. We can, however, reduce to the previous case by thinking of t' as having been determined in this way by some other person, with x_0 unknown to us, but given by $x_0 = 1/N$, where 'N is an integer chosen at random' (in the sense we discussed in Chapter 3, Section 3.2, and Chapter 4, Section 4.18).

43 Or $u_n/2 \simeq 0{\cdot}4/\sqrt{n}$: it does not make any difference whether we consider the X of the continuous case as a generalization of the length of a *string*, or of the period spent in the *lead* (i. as L or V of Chapter 8, Section 8.7).

8.9.8. If, instead, we begin by fixing a time t_0, knowing only that $Y(0) = 0$, and we consider $X = T''' - T'$, the length of the interval containing t_0 and no roots (i.e. T' = the last root of $Y(t) = 0$ with $t \leqslant t_0$, and T''' = the first root of $Y(t) = 0$ with $t \geqslant t_0$[44]), then we have the following probability distributions :

$$\text{for } T' : \ f(t) = K/\sqrt{\left[t(t_0 - t)\right]},\ F(t) = K \sin^{-1}\sqrt{(t/t_0)},\ (0 \leq t \leq t_0), \tag{8.87}$$

$$\text{for } T'' : \ f(t) = K/\sqrt{\left[t(t - t_0)\right]},\ F(t) = K \cos^{-1}\sqrt{(t_0/t)},\ (t \geq t_0), \tag{8.88}$$

$$\text{for} \quad X : \quad f(x) = K\int_{\alpha}^{t_0} \mathrm{d}t/\sqrt{\left[t(t_0 - t)(t + x)(t + x - t_0)\right]}, \qquad (x \geqslant 0), \tag{8.89}$$

where $\alpha = 0 \wedge (t_0 - x)$. Similarly, we have the result that, given t' and t'' ($t' < t''$) the probability of at least one root of $X(t)$ in the interval (t', t'') is equal to $K \cos^{-1}\sqrt{(t'/t'')}$.

The same results hold (by virtue of the usual transformations) if we think, for example, of T'' as denoting the abscissa of the maximum (or of the minimum) of $X(t)$ between 0 and t_0 (rather than the last root) and, correspondingly, of (T', T'') as the interval in which the maximum (or minimum) remains constant (i.e. T''' is the last instant up to which $X(t)$ does not exceed the maximum value attained in $(0, t_0)$; and similarly for the minimum) and so on. It is interesting to note – and this ties in with what we drew attention to in Section 9.8 as *seemingly* '*paradoxical*' – that these points (maximum, minimum, last root) are more likely to be near the end-points of the interval $(0, t_0)$ than near the centre. More precisely, as a more expressive interpretation, recall that T'' is the abscissa of a point 'chosen at random' (i.e. with uniform probability density) on the circumference of a semi-circle having $(0, t_0)$ as diameter (see Figure 8.9c).

In the case of Heads and Tails, we saw that, asymptotically, in any interval $(0, t_0)$ with $Y(0) = 0$ (or in any interval (t', t'') with $Y(t') = 0$), the proportion of time during which $Y(t)$ is positive had the arc sine distribution. This property continues to hold, exactly, for the Wiener–Lévy process.

8.9.9. The 'pathological' character of the 'small-scale' behaviour might leave one somewhat puzzled as to the possibility of interpreting the process in a constructive way. For this purpose, Lévy suggests a procedure of definition by successive approximations. It consists of subdividing the interval under consideration ($0 \leqslant t \leqslant 1$, say) into 2, 4, 8,..., 2^k,... equal parts, in determining $Y(t)$ at the division points and in taking as the kth approximation a function $Y_k(t)$, coinciding with $Y(t)$ for those t which are multiples of $1/2^k$ and linear in between them. Given $Y(0) = 0$, and $Y(1)$ determined as a random quantity with a standard normal distribution ($m = 0$, $\sigma = 1$), the intermediate points are successively determined by means of the considerations of Section 8.3.2. If t' and t'' are two consecutive multiples of $1/2^k$, and $t = (t' + t'')/2$ is the point at which $Y_{k+1}(t) = Y(t)$ is to be determined, we know that it is sufficient to add to the prevision given by $Y_k(t)$

44 If it so happened that $Y(t_0) = 0$ (and this has probability 0), we would have $T' = T''' = t_0$ and $X = 0$.

(= $[Y_k(t') + Y_k(t'')]/2 = [Y(t') + Y(t'')]/2$) a random quantity having a centred normal distribution ($m = 0$) and standard deviation given by

$$\sigma = \sqrt{\left[\left(t-t'\right)\left(t''-t\right)/\left(t''-t'\right)\right]} = 1/2^{(k+2)/2} \qquad \left(\text{see Figure 8.3}\right).$$

By bounding the probabilities of large values for these successive correction terms, we can conclude (following Lévy) that the $Y_k(t)$ converge almost certainly to a continuous $Y(t)$.

Of course, this means that we are using countable additivity. If one wishes to avoid this, all the difficulties relating to 'small-scale' behaviour could be avoided by imagining, for instance, that the process only appears to take place in continuous time but, in fact, takes place in discrete time, with time intervals $1/N$ (with N unknown and having probability 0 of being smaller than any arbitrary preassigned integer).[45]

45 Note that the same idea can be used in reverse, making any discontinuous process, for example, the Poisson process, *continuous*. It is sufficient to think of the 'jump' +1, at any instant t, as actually a continuous increase taking place in a very short time interval from t to $t + 1/N$ (with N as above; for example, one could take an increment $N\tau$ in $0 \leqslant \tau \leqslant 1/N$, or take $\sin^2\left(\frac{1}{2}\pi N\tau\right)$; one could even assume behaviour of the form $1 - e^{-N\tau}$, or $1 - e^{-N\tau} \cos N\tau$, and so on, in $0 \leqslant \tau \leqslant \infty$).
Without countable additivity, there is no unique answer to certain of the more subtle questions. Countable additivity certainly provides unique answers, but this, of course, is no reason to consider the latter as 'well founded' in any special sense.

9

An Introduction to Other Types of Stochastic Process

9.1 Markov Processes

9.1.1. The cases treated so far have been considered at some length, this being a convenient way in which to introduce various of the basic notions and most frequently used techniques. They are, however, nothing more than examples of the simplest and most special form of random process; that is, the linear form, or, more explicitly, the homogeneous process with independent increments. We now give some of the basic properties of other cases of interest, although, given the limits of the present work, the treatment will necessarily be brief.

Processes for which 'given the present, the future is independent of the past', or, alternatively, 'the future depends on the past only through the present' are called *Markov processes.* Processes with independent increments are a special case (they are even independent of the present, and the process depends on the latter only through the fact that the future increment, $Y(t) - Y(t_0)$, is added on to the present value $Y(t_0)$); the Markov property is much less restrictive.

The name derives from the fact that Markov considered this property in a particular discrete situation (involving probabilities of 'linked' events, whence *Markov chains*). To give a simple example, let us consider a function Y_n, taking only a finite number of values, 1, 2,..., r, say. For a physical interpretation, which may be more expressive, we could think of it as a 'system' which can be in any one of the r 'states', S_1, S_2,..., S_r, and which passes from one state into another in a sequence of 'steps' (including the possibility that a 'step' could result in the process remaining in the same state: recall, however, what was said in Chapter 8, 8.2.5 concerning the case $\mu_{ii} \neq 0$).

Such a system is said to be a Markov chain if, given that at time n the system is in state i, the probability that it then occupies state j at time $n+1$ is given by some value, $p_{ij}(n)$, which is independent of anything one might know about the past.

The simplest case is that of the *homogeneous* chain, for which the *transition probabilities*, p_{ij}, are also independent of the time n. These probabilities form a matrix, $\boldsymbol{P} = ||p_{ij}||$, and, with the usual definition of matrix product, its square and cube and so on give the analogous matrices of transition probabilities, $p_{ij}^{(2)}, p_{ij}^{(3)}$ and so on for passages from i to j in two steps, three steps and so on. In fact, we have

$$p_{ij}^{(2)} = \sum_h p_{ih} p_{hj} \tag{9.1}$$

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

(the sum, over h, of the probabilities of going from i to j in two steps when the intermediate state is h) and, in general,

$$p_{ij}^{(m+1)} = \sum_h p_{ih}^{(m)} p_{hj}. \tag{9.2}$$

Of course, the p_{ij} must be non-negative and, for each i, must have sum $\sum_j p_{ij} = 1$.

If the p_{ij} are all non-zero, or if this is the case for the $p_{ij}^{(m)}$ from some given m onwards, we have the so-called *ergodic* case: as m increases, the $p_{ij}^{(m)}$ tend to limit-probabilities p_j, which are independent of i. In other words, for large n, $\mathbf{P}(Y_n = j) = p_j$, independently of one's knowledge concerning the initial state i. Moreover, as n increases, the proportion of the time in which the system occupies state j during the first n steps tends stochastically to p_j (and $1/p_j$ is, in fact, the prevision of the recurrence time; i.e. the time between two successive passages through j).

If, further, we are unaware of the initial situation Y_0, and if our state of uncertainty causes us to attribute precisely these probabilities p_j as the initial, $\mathbf{P}(Y_0 = j)$, then these will remain our probabilities for the occupation of these states throughout the process, and we have what is called a *stationary* process. In fact, the vectors composed of the p_j are characterized by this property of being a fixed point (i.e. an eigenvector with unit eigenvalue) under the transformation P (and, moreover, under the stated conditions it is the unique admissible such eigenvector; i.e. with non-negative components). The ergodic property ensures that, under these conditions, we approach, asymptotically, this stationary situation. The set-up is often applied to statistical problems (involving, for example, a large number of particles or individuals etc.); then the ergodic result has a more concrete interpretation, because it implies the tendency to stationarity of the *statistical distribution.* The reader should compare this situation with those involving illegitimate applications of the ergodic 'principle', outside of the conditions under which the *theorem* holds (see Chapter 8, 8.8.4).

A similar set-up can be obtained in continuous time by assuming that the probabilities of passing from S_i to S_j in the time period from t to $t + \mathrm{d}t$ (given that we are at S_i at time t) are given by $\mu_{ij}\,\mathrm{d}t$, where the μ_{ij} may be constant, or perhaps functions of t. This case has already been considered (Chapter 8, 8.2.5) as background to our discussion of the Poisson process and we shall not add anything here to our previous discussion. As we remarked at that time, we can, without loss of generality, take $\mu_{ij} = 0$ for $i = j$ (and, except for special cases, it is usually convenient to do so).

9.1.2. Within the framework of the very simple cases that formed the basis of our previous discussion, we outlined several of the main problems and features of interest that arise with these processes. The same kinds of problems and features have been studied for general Markov processes and, without going into the details, we shall consider a few of these in order to make some appropriate comments.

The kind of relation which we have encountered in the simple form $p_{ij}^{(2)} = \sum_h p_{ih} p_{hj}$ is typical of the Markov set-up (even in continuous time and with continuous state space). Given that we start from some point P_0 at time t_0, the probability of being in a neighbourhood of a point P_1 at time t_1 satisfies a relation involving the sum (or infinite series, or integral, as the case may be) of the probabilities of getting there by passing

through the various possible points P at some arbitrary, intermediate time t ($t_0 < t < t_1$). These probabilities are evaluated as the product of the probabilities of the two passages: from P_0 to (a neighbourhood of) P in $[t_0, t]$, and then from P to (a neighbourhood of) P_1 in $[t, t_1]$, the latter probability being independent of P_0. This is the probabilistic version of 'Huyghens' principle', by which, in the deterministic case, one regards the evolution of a system in the period from t_0 to t_1 as being the result of what happens between t_0 and t, followed by what happens from t to t_1, starting from the situation reached at t with no need to recall the past. In our case, the same thing applies, not to the evolution of the system, but to the evolution of the probability distribution on the basis of which we foresee the evolution of the system.

In both cases (Huyghens and Markov), these processes are sometimes referred to as *nonhereditary* (in contrast to *hereditary* phenomena, whose evolution is influenced by the past). Examples are provided by the phenomenon of hysteresis, Volterra integral equations and so on. One should note, however, that, in these respects, the distinction between 'present' and 'past' is something of a convention. One often believes that a (deterministic) prediction or a (probabilistic) prevision would be determined by the present if only some (unattainable) data or measurements were known. To compensate for their unavailability, one makes use of available data relating to the past (for example, in the case of hysteresis, the characteristics of the present situation are deduced from the history of the magnetic field which has produced them, since it is impossible to explore the state of magnetization at each point of a body). At an even more basic level, it can happen that for some problems 'the present' can be regarded as the position of a particle (or a body etc.), whereas for others we need, in addition, to know the velocity (or the last movement). This is also true in the probabilistic case and we can consider a *second-order* Markov chain as one in which the probabilities of the possible values for $Y(n+1)$ depend both on the value of $Y(n)$ and of $Y(n-1)$ (but on no others; one could, however, extend the notion and consider chains of arbitrary order). In fact, we could reduce this directly to the first-order case by defining 'the present' at time n to consist of the pair

$$\Big(Y\big(n-1\big), Y\big(n\big)\Big).$$

In other words, we redefine the 'states' to be the r^2 pairs (i, j), with the obvious restriction that from a state (i, j) one can only move, in one step, to one of the r states of the form (j, h).

9.1.3. Although it may seem a 'natural' condition, we are not claiming that the Markov property holds in all 'nonpathological' cases, nor even for the simplest, standard processes. Simple counterexamples that are of practical interest arise in connection with the Poisson process. For example, let $Y(t)$ be the number of telephone conversations in progress on some telephone system at time t and let us assume that $N(t)$, the number of conversations which began between 0 and t, has a Poisson distribution, and that the length of any conversation is a random quantity having the same distribution as all the others, and stochastically independent of them. If this distribution is exponential, the process is Markovian (because every conversation in progress then has the same probability, $\lambda \mathrm{d}t$, of terminating within an infinitesimal time $\mathrm{d}t$, whatever its duration so far) but, in every other case, knowledge of the duration of the conversation so far modifies the prevision. In other examples of this kind, as here, 'age', or something similar, plays a fundamental rôle. A similar kind of example is that where the 'cumulative

effects' have an influence; the prevision at t_0 depends not only on $Y(t_0)$ but also on the sum (or integral, in the continuous case) of the values $Y(t)$ between 0 and t_0. Of course, if the ages or the cumulative values were included as part of the definition of 'the present' (and were observable, or somehow available) then the process, appropriately extended to include these other variables, would turn out to be Markovian.

9.2 Stationary Processes

9.2.1. We have already given the basic idea of a *stationary* process. We discussed it in relation to a Markov process, but this is not a necessary condition for stationarity. A sufficient condition is that the probabilities are invariant with respect to a translation along the time axis. For example, it is sufficient that the probabilities of $Y(t)$, $Y(t+t_1)$,..., $Y(t+t_k)$ satisfying the inequalities $y_i' \leqslant Y(t+t_i) \leqslant y_i''$ are independent of t. The above example of a telephone system (along with similar examples) gives a stationary process if we assume either that the system has been in operation for an infinitely long time or that Y is unknown and that we attribute to the values that it can assume at each instant those probabilities corresponding to the assumption of an infinitely distant beginning. We note that the process is Markovian or non-Markovian according to whether the distributions of the lengths of conversations are or are not independent exponential distributions.

If $Y(t)$ is a stationary process, then the definition implies, in particular, that the distribution of $Y(t)$ does not depend on t, and so neither does the prevision (if it exists) $\mathbf{P}(Y(t)) = m$, nor the variance $\mathbf{P}(|Y(t)|^2) = \sigma^2$ (we assume, for convenience, and without loss of generality, that $m = 0$ and $\sigma = 1$). The same holds for all other moments and parameters of the distribution. If we consider two distinct times t' and t'', the pair of values $Y(t')$ and $Y(t'')$ has a distribution depending only on the difference $t'' - t' = u$,[1] and, in particular, all the quantities defined in terms of this distribution depend only on u; above all, this applies to the *correlation*

$$\phi(u) = r\left(Y(t'), Y(t'')\right) = \mathbf{P}\left(Y(t')Y^*(t'')\right).^2 \tag{9.3}$$

This correlation – usually referred to as the *autocorrelation* function – characterizes the process so far as second-order properties are concerned (in the sense that it enables one to determine $\mathbf{P}(X)$ for all $X = \sum a_{ij} Y_i Y_j^*$; i.e. for functions of the second degree in the values $Y_i = Y(t_i)$ of $Y(t)$ at any number of arbitrary time points t_i). If we have $m \neq 0$ and $\sigma \neq 1$, we can get back to the original process from the standardized case by noting that the former is equal to $m + \sigma Y(t)$. Similar conclusions hold in the nonstationary case, also, provided that $\mathbf{P}(Y(t))$ is constant and

$$\mathbf{P}\left(Y(t')Y^*(t'')\right) = \Gamma\left(t', t''\right) \tag{9.4}$$

---

1 The choice of u is a deliberate attempt to exploit an analogy that makes it convenient to use the same notation, $\phi(u)$, for both the autocorrelation and the characteristic function (see the next section).

2 The asterisk denotes 'complex conjugate'. For the present, it is superfluous, as we are only considering real functions; shortly, however, we shall need the extension to the complex field.

depends only on the difference between t' and t''. Putting $t' = t''$ gives the second moment and if this is bounded so is Γ, and the process is called 'second-order stationary'.

9.2.2. In dealing with this topic it is convenient to allow $Y(t)$ to be complex (for the same reasons for which it is convenient to represent harmonic oscillations with e^{it} rather than sines and cosines). The product $Y(t')Y(t'')$ therefore has to be replaced by the Hermitian product; that is by $Y(t')Y^*(t'')$ (as we already indicated when we defined ϕ and Γ). This implies that

$$\Gamma\left(t'', t'\right) = \Gamma^*\left(t', t''\right),$$

and, in particular, that $\phi(-u) = \phi^*(u)$. The latter is the more important because it relates directly to the stationary case that we are discussing. Moreover, the real part of $\phi(u)$ is continuous if (and only if) the process is 'mean-square continuous' (a stationary process enjoying this property is known as a *Khintchin process*). This property requires, in the notation of Chapter 6, 6.8.3, that $Y(t) \xrightarrow{\bullet} Y(t_0)$ as $t \to t_0$, but – and one should be clear about this – it says nothing about the continuity of the function $Y(t)$. We require that the prevision of $[Y(t) - Y(t_0)]^2$ tends to 0. This happens, for example, for a Poisson process, or variants thereof, even for generalized Poisson processes (these only change through discontinuities, which, in the latter case, are everywhere dense), provided the standard deviation is finite (in this case, in fact, $\mathbf{P}(Y(t) - Y(t_0))^2 = K|t - t_0| \to 0$).

Under these conditions, it can be shown that the class of possible correlation functions coincides with the class of characteristic functions (and, of course, in the case of $Y(t)$ with even-valued, and hence real-valued, characteristic functions, we have

$$\phi\left(-u\right) = \phi^*\left(u\right) = \phi\left(u\right),$$

which correspond, as characteristic functions, to symmetric distributions $F(-x) = 1 - F(x)$). In any case, the distribution F has an important significance so far as the process is concerned, not only from a mathematical point of view but also practically, in all applications, especially to problems in physics, where it has a connection with energy. It gives, in fact, the *spectral function* of the process: that is $F(\omega_2) - F(\omega_1)$ is the prevision of the energy corresponding to the frequencies in the interval $\omega_1 \leqslant \omega \leqslant \omega_2$. Expressed in an informal manner, the actual meaning of this in relation to the random function $Y(t)$ defined by the process is the following: let $U(\omega)$ (in general, complex) be the function expressing $Y(t)$ as a mixture of harmonic components (i.e. as a Fourier–Stieltjes transform)

$$Y\left(t\right) = \int_{-\infty}^{\infty} e^{i\omega t} dU\left(\omega\right). \tag{9.5}$$

The prevision of the energy corresponding to an individual $d\omega$ is

$$dF\left(\omega\right) = \mathbf{P}\left(\left|dU\left(\omega\right)\right|^2\right), \tag{9.6}$$

and in terms of F we obtain the correlation function

$$\phi\left(u\right) = \int_{-\infty}^{\infty} e^{i\omega u} dF\left(\omega\right), \tag{9.7}$$

which is therefore the characteristic function of the energy distribution.

The spectrum F could contain both concentrated masses (jumps of F)

$$U_k = F(\omega_k + 0) - F(\omega_k - 0),$$

corresponding to 'lines' ω_k and diffused masses (segments where F is increasing and continuous). To make things as clear as possible, we repeat and extend the previous discussion in the simpler case where we just have concentrated masses U_k, corresponding to a set of particular frequencies ω_k. In this case, we have

$$Y(t) = \sum_k U_k e^{-i\omega_k t}, \tag{9.5'}$$

and we can deduce that

$$\begin{aligned} U_k &= \lim_{a\to\infty} \frac{1}{2a} \int_{-a}^{a} e^{-i\omega_k t} \sum_h U_h e^{i\omega_h t} dt \\ &= \lim_{a\to\infty} \frac{1}{2a} \int_{-a}^{a} e^{-i\omega_k t} Y(t) dt. \end{aligned} \tag{9.8}$$

The U_k are, therefore, random quantities that depend on the global behaviour of $Y(t)$; conversely, knowledge of these random quantities determines $Y(t)$ in the way we have indicated. To give the probability distribution for all the U_k is an indirect way of giving all the probabilities of the process leading to $Y(t)$. From the energy viewpoint, we could say that the energy for the frequency ω_k is $|U_k|^2$ with prevision $\mathbf{P}(|U_k|^2)$; the U_k are uncorrelated (i.e. 'orthogonal') in the sense that $\mathbf{P}\left(U_h U_k^*\right) = 0$ $(h \neq k)$, and the total energy for frequencies up to and including ω is given by

$$F(\omega) = \sum_k \mathbf{P}\left(|U_k|^2\right) \tag{9.6'}$$

(the sum being taken over all the k for which $\omega_k \leqslant \omega$).

Because we have standardized the process ($\sigma = 1$), the total energy equals 1. The correlation function is given by

$$\phi(u) = \sum_k \mathbf{P}\left(|U_k|^2\right) e^{i\omega_k u}. \tag{9.9}$$

Cramèr and Loève have proved that in the case we have considered (with discrete spectrum) the U_k are mutually orthogonal and that this also holds in the general case (either second-order stationary or Khintchin processes) for the $dU(\omega)$ for disjoint intervals $d_1\omega$ and $d_2\omega$;

$$\mathbf{P}\left(d_1 U . d_2 U^*\right) = 0.$$

Conversely, all such processes can be obtained in this way (a result also proved by Cramèr and Loève).

9.2.3. The concepts and techniques that we have discussed are not only applicable to problems in physics – from which we have borrowed the particularly expressive form

of terminology – but also to problems in other fields, such as statistics ('time-series' analysis and so on). However, this was a good place to give an outline of these ideas; it is useful to be able to 'see' the various problems concerning Fourier transforms and their mathematical properties in terms of some concrete framework. The two applications we have encountered are, in some sense, mutual inverses one of the other. In the case of the characteristic function, the concrete datum, or, at any rate, the most immediate, was the distribution, and the transform mainly provided a 'useful image' of it; in the case we have just dealt with, the function $Y(t)$ and the autocorrelation function are more concrete, and the corresponding distributions U and F are the 'images' in a certain wave interpretation.

10

Problems in Higher Dimensions

10.1 Introduction

10.1.1. It might be argued that every problem could, or even should, be put in a multidimensional framework; indeed, we have seen this over and over again throughout our treatment so far. The subject matter of this chapter is not really new, therefore, and we shall merely emphasize those features and problems which particularly relate to the multi-dimensional nature of certain distributions.

In Chapter 6, 6.9.1, we dealt with the essential points concerning the representation of a distribution over an r-dimensional Cartesian space, either by means of the distribution function

$$\begin{aligned} F\left(x_1, x_2, \ldots, x_r\right) &= \mathbf{P}\left[\left(X_1 \leqslant x_1\right)\left(X_2 \leqslant x_2\right)\ldots\left(X_r \leqslant x_r\right)\right] \\ &= \mathbf{P}\left[\prod_i \left(X_i \leqslant x_i\right)\right], \end{aligned} \tag{10.1}$$

or, if it exists, by means of the density

$$f\left(x_1, x_2, \ldots, x_r\right) = \partial^r F / \partial x_1 \partial x_2 \ldots \partial x_r . \tag{10.2}$$

In addition, we can state that a necessary and sufficient condition for a function $F(x_1, x_2, \ldots, x_r)$ to be a distribution function is that f never be non-negative, or, should f not exist, that the expression for which it would be the limit is non-negative. The latter is the probability of the rectangular prism $\left(x_i' < X_i \leqslant x_i''\right)\left(i = 1,2,\ldots,r\right)$ given by

$$\mathbf{P}\left[\prod_i \left(x_i' < X_i \leqslant x_i''\right)\right] = \sum \pm F\left(x_1, x_2, \ldots, x_r\right), \tag{10.3}$$

the sum being taken over the 2^r vertices corresponding to all possible assignments of $x_i = x_i'$ or $x_i = x_i''$, with a + or – sign according to whether these are an even or odd number of x' (for the case $r = 2$, see Figure 6.5 in Chapter 6, 6.9.1, together with the intuitive explanation that accompanied it).

In order to 'see' the meaning of this condition (which is a generalization of the nondecreasing property of the one-dimensional F), it is useful to think of a mass c placed at

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

some given point $P_0 = \left(x_1^0, x_2^0, \ldots, x_r^0\right)$ as giving rise to a 'step' of height c in that orthant[1] of the r-dimensional space of the points whose coordinates are greater than the corresponding x_i^0 (in the plane, this would be the NE quadrant). The function F is given by the superposition of such steps (or as a limit case).

The disadvantage of this is that the function F depends on the coordinate system (often, however, the problem itself has arisen in connection with r given random quantities X_i). A less arbitrary – but less useful – approach would be to assign probabilities over each half-plane (i.e. to assign $F(y)$ for each linear combination $Y = \sum_i a_i X_i$). The justification for this is straightforward, although somewhat indirect, and follows from the fact that this serves to determine the characteristic function, which, in turn, determines the distribution (as we shall see in the next section).

10.1.2. The characteristic function for an r-dimensional distribution of $X_1, X_2, \ldots, X_r$ is a function of r variables, $u_1, u_2, \ldots, u_r$, defined in a completely analogous way to that in the one-dimensional case:

$$\phi\left(u_1, u_2, \ldots, u_r\right) = \mathbf{P}\left(\mathrm{e}^{\mathrm{i}\left(u_1 X_1 + u_2 X_2 + \ldots + u_r X_r\right)}\right) = \mathbf{P}\left(\mathrm{e}^{\mathrm{i}\mathbf{u}\times\mathbf{X}}\right). \tag{10.4}$$

The vector form is probably the clearer, with vectors $\mathbf{X}$ and $\mathbf{u}$ whose components are the X_i and u_i, respectively ($\mathbf{u}\times$ can, if we so wish, be regarded as a vector in the dual space).

For the cases $r = 2$ and $r = 3$, it is more convenient to avoid the use of subscripts and to write $uX + vY$, $uX + vY + wZ$, respectively (the standard notation for Plückerian coordinates).

The properties of $\phi(u_1, u_2, \ldots, u_r) = \phi(\mathbf{u})$ are (as is fairly obvious) the same as in the one-dimensional case. The inversion formula is also the same: for the case $r = 2$, for example, if the density exists and is bounded, it is given by

$$f\left(x, y\right) = \frac{1}{\left(2\pi\right)^2} \int\int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}(ux+vy)} \phi\left(u, v\right) \mathrm{d}u\,\mathrm{d}v. \tag{10.5}$$

If, in addition, the X_h are independent, we have

$$F\left(x_1, x_2, \ldots, x_r\right) = F_1\left(x_1\right) F_2\left(x_2\right) \ldots F_r\left(x_r\right), \tag{10.6}$$

$$\phi\left(u_1, u_2, \ldots, u_r\right) = \phi_1\left(u_1\right) \phi_2\left(u_2\right) \ldots \phi_r\left(u_r\right), \tag{10.7}$$

as well as the converse; that is factorization implies stochastic independence.

10.1.3. A number of problems in higher dimensions can be dealt with formally as though they were one-dimensional problems by means of matrix and vector notation. For example, sums of random vectors have the same properties as sums of random quantities. In particular, if the vector summands (each with prevision zero) all have the same distribution and finite variances, then the sum-vector of n of them, divided by $\sqrt{n}$, has, asymptotically, a normal distribution having the same variances and covariances.

A frequently used and very expressive interpretation is that in terms of a 'random walk' in r-dimensional space, regarded as a random process in discrete time (as an aid

1 Orthant is the r-dimensional analogue of half-line ($r = 1$) and quadrant ($r = 2$).

to intuition, we shall mainly deal with the cases $r = 2$ and $r = 3$, corresponding to the plane and ordinary space); a step is taken after each unit of time and at each step we obtain a random vector (always with the same distribution and stochastically independent). The simplest example is obtained, for example, by simultaneously studying the gain of two (or three) gamblers who bet independently on a sequence of tosses at Heads and Tails (again ±1 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$ at each toss). This results in a zigzag path (in the plane, each step from (x_n, y_n) to (x_{n+1}, y_{n+1}) is the diagonal of some square in the integer lattice; the same holds in three dimensions with the diagonals of cubes). If (X_n, Y_n) is the 'position after n tosses', then, as n increases, it can be shown that this has, asymptotically a normal distribution with circular symmetry and standard deviation $\sqrt{n}$ in all directions (and the same holds for the position (X_n, Y_n, Z_n) in three dimensions).

10.1.4. The following is a simple and instructive argument that can be applied to the present case. The probability of a return to the origin after n tosses in the one-dimensional case is given by $u_n \simeq 0 \cdot 8/\sqrt{n}$ for n even, 0 for n odd. In the case of the plane ($X_n = Y_n = 0$), or ordinary space

$$\left(X_n = Y_n = Z_n = 0\right),$$

the respective probabilities are therefore given by $u_n^2 \simeq 0.64/n$ and $u_n^3 = 0 \cdot 51/\sqrt{n^3}$: in the general case, we have $u_n^r = K/n^{r/2}$. We observe immediately that, in prevision, the number of returns to the origin is infinite in the plane ($\sum n^{-1}$ diverges) but is finite in three dimensions ($\sum n^{-r/2}$ converges for $r \geqslant 3$). It follows that the return to the origin is practically certain ($p = 1$) for $r = 1$ and $r = 2$ but not for $r \geqslant 3$ (where $p = a/(1 + a)$, with $a = \sum n^{-r/2}$; for $r = 3$, for example, $a \simeq 0 \cdot 53$ and $p \simeq 0 \cdot 35$[2]).

The conclusion concerning the limit distribution (normal, with rotational symmetry and dimensions increasing like $\sqrt{n}$) holds in the general case, also, provided the distribution of every individual step has the same variance in all directions (i.e. equal variances and zero correlation for any two orthogonal directions). Without these conditions, we would have 'ellipsoidal contours' instead of spheres (but the latter case can be reduced to the former by making appropriate changes of scale along the axes of the ellipsoids).

10.2 Second-Order Characteristics and the Normal Distribution

10.2.1. To illustrate the use of vector and matrix notation, we shall re-examine certain expressions that we have already encountered in the context of the multivariate normal distribution, pointing out the form that certain properties now take.

The notation we shall introduce will enable us to interpret and understand our formulae in several alternative ways: either in the rather formalistic spirit that derives from algebraic-type theories (vectors and matrices thought of as rows, or columns, or arrays of numbers) or in the geometric, functional analytic spirit.

---

2 If p is the probability of (at least) one return to the origin, $(1 - p)p^h$ is the probability of exactly h returns to the origin, and the prevision of the number of returns is given by

$$a = \sum hp^h(1 - p) = p/(1 - p)$$

Vectors will be written boldface: for example, $\mathbf{x}$ (or $\mathbf{X}$, if we are dealing with a random vector). Given r linearly independent vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ in S_r, $\mathbf{x}$ can be written (in one and only one way) as a linear combination of them; $\mathbf{x} = \sum x_h \mathbf{u}_h$. We may sometimes write $\mathbf{x} = (x_1, x_2, \ldots, x_r)$, but this is simply a convention and leaves it to be understood (and never forgotten) that the components do not directly relate to the intrinsic meaning of the vector, but only acquire their meaning through the introduction of some arbitrary basis, which can be changed at any time, the choice being simply a matter of convenience (this conflicts somewhat with the algebraic viewpoint). For a random $\mathbf{X}$, we shall write $\mathbf{X} = \sum \boldsymbol{X}_h \mathbf{u}_h = (X_1, X_2, \ldots, X_r)$. The linear functional on the vectors of the space S_r themselves form an r-dimensional space, the *dual* space, which we shall denote by S_r^*.

10.2.2. If we introduce a *metric* into the space S_r (i.e. a *scalar product*, which maps each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ to a scalar, $\mathbf{x} \times \mathbf{y} = \mathbf{y} \times \mathbf{x}$, and is linear for each vector, and such that $\mathbf{x} \times \mathbf{x} > 0$ for all $\mathbf{x}$ other than the zero vector), then each dual vector can be expressed as a vector of the original space with the scalar product sign following it. In other words, if $f(\mathbf{x})$ is a scalar depending linearly on $\mathbf{x}$, then there exists a vector $\mathbf{a}$ such that $f(\mathbf{x}) = \mathbf{a} \times \mathbf{x}$, and $f(\cdot)$ can be written as $\mathbf{a}\times$. Given a metric in S_r, it makes sense to define the norm of a vector, $|\mathbf{x}| = \sqrt{\mathbf{x} \times \mathbf{x}}$, and the orthogonality of two vectors, $\mathbf{x} \times \mathbf{y} = 0$. It then becomes convenient to choose the basis to be an orthogonal set of $\mathbf{u}_h$ with unit norms, in which case we denote them by $\mathbf{i}_h$:

$$\left(\mathbf{i}_h \times \mathbf{i}_k = \left(h = k\right); \text{ i.e. 1 or 0, according as } h = k \text{ or not}\right).$$

The scalar product then has a simple representation in terms of the components: $\mathbf{x} \times \mathbf{y} = \sum x_h y_h$ (and $|\mathbf{x}| = \sqrt{(\sum^{x_h^2})}$) We shall write $\mathbf{a}^*$ instead of $\mathbf{a}\times$, and $\mathbf{a}^*$ is then interpreted as the 'dual of $\mathbf{a}$' (some authors write $\mathbf{a}^{\mathrm{T}}$, where the superscript denotes 'transpose'; others use $\mathbf{a}_{-1}$; and so on). These alternative notations relate to the interpretation of the vectors in the two spaces as 'column vectors' or 'row vectors', respectively (i.e. matrices with 1 column and r rows, or 1 row and r columns).

From the formal, algebraic point of view, the matrices are also considered simply as arrays of numbers (r rows and s columns). From the geometric or functional analytic point of view, they are linear transformations between some S_r and some S_s. In our particular case, we shall only be considering square matrices.

If A is a matrix (or, better, a *linear transformation*), we have

$$\mathbf{y} = A\mathbf{x}, \quad \text{with } A\left(\mathbf{x}_1 + \mathbf{x}_2\right) = A\mathbf{x}_1 + A\mathbf{x}_2,\ A\left(c\mathbf{x}\right) = cA\mathbf{x}. \tag{10.8}$$

In terms of components, if $\mathbf{x} = \sum x_h \mathbf{i}_h, A\mathbf{x} = \sum x_h A\mathbf{i}_h$, we have

$$A\mathbf{i}_h = a_{h1}\mathbf{i}_1 + a_{h2}\mathbf{i}_2 + \ldots + a_{hr}\mathbf{i}_r, \text{ and}$$

$$A\mathbf{x} = \sum_h x_h \sum_k a_{hk}\mathbf{i}_k = \sum_k \left(\sum_h a_{hk} x_h\right)\mathbf{i}_k : \tag{10.8′}$$

in other words, the components of $\mathbf{y} = A\mathbf{x}$ are given by $y_k = \sum_h a_{hk} x_h$. The linear transformation A can therefore be represented (in the given reference system) by means of the r^2 coefficients a_{hk} (which, in the array, corresponds to the hth row, kth column).

10.2.3. We are particularly interested in those linear transformations (or matrices) which, with respect to the metric under consideration, are symmetric and positive; that is they correspond to 'positive-definite quadratic forms':

$$A\mathbf{x}\times\mathbf{y} = A\mathbf{y}\times\mathbf{x},\ \ A\mathbf{x}\times\mathbf{x} > 0 \text{ provided } \mathbf{x}\neq 0.$$

If Q denotes such a linear transformation, we shall make the convention that Q will also be used to denote the matrix and the quadratic form. We can write, therefore,

$$Q\{\mathbf{x}\} = Q\mathbf{x}\times\mathbf{x} = Q\mathbf{x}^{*}\mathbf{x} = \mathbf{x}^{*}Q\mathbf{x} = \mathbf{x}^{\mathrm{T}}Q\mathbf{x} = \sum_{hk} q_{hk}x_h x_k \ (q_{hk} = q_{kh}) \tag{10.9}$$

(where the symbols are to be interpreted in an appropriate way). Everything is straightforward, except that, in order to conform with the standard conventions of matrix manipulation, we would need to write $\mathbf{x}A$ instead of $A\mathbf{x}$, $\mathbf{y}\mathbf{x}^{\mathrm{T}}$ instead of $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ (corresponding to $\mathbf{x}^{*}\mathbf{y}$ or $\mathbf{x}\times\mathbf{y}$), and, therefore, $\mathbf{x}Q\mathbf{x}^{\mathrm{T}}$ instead of $Q\mathbf{x}^{*}\mathbf{x}$. All vectors are to be understood as row vectors, except when they have a 'transpose' superscript, which transforms them into column vectors (dual vectors; i.e. of the form $\mathbf{a}\times$, but as operators on the right). Note, therefore, that while $\mathbf{x}\mathbf{y}^{\mathrm{T}}$ means $\mathbf{y}\times\mathbf{x}$, $\mathbf{y}^{\mathrm{T}}\mathbf{x}$ means $\mathbf{x}.\mathbf{y}\times$; that is it represents the transformation $\boldsymbol{A}$ which takes every vector $\mathbf{z}$ to $\boldsymbol{A}\mathbf{z} = \mathbf{x}.(\mathbf{y}\times\mathbf{z})$ (the transformation of rank 1 which transforms all the vectors of S_r into vectors parallel to a particular vector; $\mathbf{x}$ in our case); the entries of the matrix $\boldsymbol{A}$ are given by $a_{hk} = x_k y_h$.[3] Observe, in particular, that $\mathbf{x}.\mathbf{x}\times$, or $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ (such that $\boldsymbol{A}\mathbf{z} = \mathbf{x}.(\mathbf{x}\times\mathbf{z})$), represents the vector which is the projection of $\mathbf{z}$ in the direction of $\mathbf{x}$ (if $\mathbf{x}$ is a unit vector; otherwise, it is multiplied by $\mathbf{x}^2$, which we write instead of $|\mathbf{x}|^2$, i.e. $\mathbf{x}\times\mathbf{x}$).

10.2.4. The covariance matrix – defined in Chapter 4, 4.17.5, for random variables X_h with $\mathbf{P}(X_h) = 0$, by $\sigma_{hk} = \mathbf{P}(X_h X_k)$ – can be defined in this set-up as Var($\mathbf{X}$), or simply $V(\mathbf{X})$, by setting, for $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_r)$,

$$V(\mathbf{X}) = \mathbf{P}\left(\mathbf{X}.\mathbf{X}\times\right) = \mathbf{P}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right):$$

in other words, as the linear transformation which gives, for each vector $\mathbf{u}$,

$$V(\mathbf{X})\mathbf{u} = \mathbf{P}\left(\mathbf{X}.\left(\mathbf{X}\times\mathbf{u}\right)\right). \tag{10.10}$$

Since

$$V(X)\mathbf{u}\times\mathbf{v} = \mathbf{P}\left(\left(\mathbf{X}\times\mathbf{u}\right)\left(\mathbf{X}\times\mathbf{v}\right)\right) = V(\mathbf{X})\mathbf{v}\times\mathbf{u},$$

the linear transformation is symmetric, so we can find an r-tuple of orthogonal directions which are mapped to themselves (i.e. there exist eigenvectors $\mathbf{v}_h$ and eigenvalues λ_h such that $V(\boldsymbol{X})\mathbf{v}_h = \lambda_h\mathbf{v}_h$); the transformation is also positive ($V(\boldsymbol{X})\mathbf{u}\times\mathbf{u} = \mathbf{P}(\mathbf{X}\times\mathbf{u})^2$), and so $\lambda_h > 0$.

When we are referring to a fixed $\mathbf{X}$, and there is no danger of ambiguity, we shall simply write $\boldsymbol{V}$ in place of $\boldsymbol{V}(\mathbf{X})$.

---

3 This follows, even without taking into account the geometrical meaning of x_x and y_h, from the fact that the characteristic of the matrix must be 1 (rows and columns are proportional).

We have already seen (in Chapter 4, 4.17.5, and in Chapter 7, 7.6.7) that the normal distribution, in whatever number of dimensions, is characterized by its covariance matrix and that such a matrix (i.e. symmetric and positive definite) characterizes a unique normal distribution (where throughout we are assuming distributions to be centred at zero). At the point $0+\mathbf{x}$ the density has the form

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_r) = K\mathrm{e}^{-\frac{1}{2}Q\{\mathbf{x}\}},\ K = 1/\sqrt{\left[(2\pi)^r \det Q\right]}. \tag{10.11}$$

The relationship of Q and V is given by $V = Q^{-1}$ (and, conversely, $Q = V^{-1}$), by virtue of the fact that the eigenvalues are the variances, σ_h^2, for V, but their inverses, σ_h^{-2}, for Q.

For these reasons, we again get involved with the *ellipsoid of covariance* (or *of inertia*) and the *ellipsoid of concentration*, which we first came across in Chapter 4, 4.17.6, and which we are forced to consider further. What we said in Chapter 7, 7.6.7, concerning the affine properties (for which it is sufficient to consider the case of spherical symmetry) still holds, whereas any consideration of the ellipsoids only makes sense, and has any use, if it is necessary, or appropriate, to base oneself upon a preassigned metric. (This would be the case, for example, were we dealing with a problem in real, physical space, or if a number of problems, each of which separately would require a different metric for convenience, were considered simultaneously.)

In any case, we lose nothing in the way of generality, and we gain a great deal in terms of simplicity and understanding, if, in order to study this problem, we take the principal axes of inertia as our reference system. In other words, we take as our unit vectors $\mathbf{i}_h$ the eigenvectors of Q and V (necessarily orthogonal)[4], whose respective eigenvalues are the variances σ_h^2 of V and the reciprocals ('weights') σ_h^{-2} of Q.

We obtain, therefore,

$$Q\mathbf{i}_h = \sigma_h^{-2}\mathbf{i}_h,\ Q\mathbf{x} = \sum_h x_h\sigma_h^{-2}\mathbf{i}_h, \tag{10.12}$$

$$Q\{\mathbf{x}\} = Q\mathbf{x}\times\mathbf{x} = \sum_h \left(\sigma_h^{-2}x_h\right)x_h = \sum_h \sigma_h^{-2}x_h^2 = \sum_h \left(x_h/\sigma_h\right)^2; \tag{10.13}$$

$$V\mathbf{i}_h = \sigma_h^2\mathbf{i}_h,\quad V\mathbf{u} = \sum_h u_h\sigma_h^2\mathbf{i}_h, \tag{10.14}$$

$$V\{\mathbf{u}\} = V\mathbf{u}\times\mathbf{u} = \sum_h \left(\sigma_h^2 u_h\right)u_h = \sum_h \sigma_h^2 u_h^2 = \sum_h \left(\sigma_h u_h\right)^2. \tag{10.15}$$

As we already know, $Q\{\mathbf{x}\}$ is useful when it comes to expressing the density (by means of equation 10.11), which, in the present reference system, becomes (there being no cross-product terms)

$$f(\mathbf{x}) = K\mathrm{e}^{-\frac{1}{2}Q\{\mathbf{x}\}} = K\exp\left[-\frac{1}{2}\sum_h \left(x_h/\sigma_h\right)^2\right] = K\prod_h \exp\left[-\frac{1}{2}\left(x_h/\sigma_h\right)^2\right]. \tag{10.16}$$

This shows (as was obvious anyway) that, in this reference system, the components X_h of $\mathbf{X}$ are stochastically independent (the density is a product of factors each of which

4 Apart from irrelevant ambiguities in the case of multiple eigenvalues.

is a function of only one x_h). But this implies that the same factorization holds for the characteristic function,

$$\phi(\mathbf{u}) = \prod_h \exp\left[-\frac{1}{2}(\sigma_h u_h)^2\right] = \exp\left[-\frac{1}{2}\sum_h (\sigma_h u_h)^2\right] = \mathrm{e}^{-\frac{1}{2}V\{\mathbf{u}\}}, \tag{10.17}$$

and we therefore see the complementary rôle played by $V = Q^{-1}$ in defining the characteristic function.

The two ellipsoids are given by

$$V\{\mathbf{u}\} = 1\big(\textit{covariance} \text{ or } \textit{inertia}; \text{semi-axes } 1/\sigma_h\big), \quad \text{and}$$

$$Q\{\mathbf{x}\} = 1\big(\textit{concentration}; \text{ semi-axes } \sigma_h\big).$$

The choice of the different variables **u** and **x** for V and Q is deliberate, and in explaining this choice we will be led to a comparison of the two ellipsoids. The **x** on which Q operates are the actual vectors of the space over which the distribution is defined (the ambit $\mathscr{A}$; e.g. physical space): the **u** on which V operates are essentially the dual vectors (even though, given the introduction of the metric, the two spaces are superposed). This supports the idea that the ellipsoid of concentration is more directly meaningful, as was confirmed, in part, by what we established in Chapter 4, 4.7.6. We must now, as we then promised, consider this further, basing ourselves on the representation in terms of the appropriate normal distribution; that is the distribution with the most frequently occurring and stable form having the same previsions and covariances (in mechanical terms, barycentre and kernel of inertia).

As we have seen, the ellipsoids $Q =$ constant are the surfaces on which the density, f, is constant. The special case $Q = 1$ (which gives, therefore, $f = K\mathrm{e}^{-\frac{1}{2}}$, which is $0{\cdot}606$ of the maximum at the origin) enjoys a property that justifies one in singling out, and defining as the *body* or *kernel* of the distribution, that part of it contained in $Q \leqslant 1$ (that part corresponding to $Q \leqslant 1$ might be referred to as the *tail*, or *shell*, but no appropriate term seems to exist). The meaning is clearest in one dimension: the kernel is the portion of the distribution with convex density lying between the points of inflexion – see Figure 7.6 in Chapter 7, 7.6.6 – and the tail consists of the two outside portions with concave density; that is tapering away. The same thing applies in the general case, however: *inside* $Q \leqslant 1$ the density is convex (with the same meaning: for each point $\lambda A + (1-\lambda)B$, $0 < \lambda < 1$, in the segment between A and B, the density has a greater value than the linear interpolation $\lambda f(A) + (1-\lambda)f(B)$); *outside*, however, that is for $Q \geqslant 1$, in the direction of a radius emanating from the origin the behaviour is concave (we are back in the one-dimensional case), and convex in all directions which are conjugate with respect to Q.

10.3 Some Particular Distributions: The Discrete Case

10.3.1. We shall now look at a few specific problems in more detail, and we begin with those involving discrete distributions.

Many of the problems we have considered for ordinary events can be extended in an obvious manner to the case of multi-events: instead of a coin, which only has two faces,

we could consider a die, which has six faces; instead of an urn with black and white balls, we could have an urn containing balls of r different colours; instead of games that can only result in either victory or defeat, we could consider those in which a draw is also possible, or we could even distinguish a whole range of results (for example, the actual scores, 3–1, 2–2, 0–1 etc., as in football), and so on.

In all these cases, by making various assumptions, there are a whole range of problems that can be considered. In particular, one can try to calculate the probabilities of the r possibilities 1, 2,..., k,..., r occurring n_1, n_2,..., n_k,..., n_r times, respectively. This same question can be formulated along different lines, clearly equivalent, but seemingly different at first sight. For example, we might ask how many objects will be given to each of r individuals as the result of some given method of selection (like giving an object to individual k whenever a certain outcome occurs). If the 'objects' are 'particles', and instead of individuals we think of 'physical states', or 'cells' corresponding to them, the different distributions will correspond to different 'macroscopic states'.

10.3.2. The following examples are of this kind and in order to make them seem more intuitive we shall present them as far as possible in terms of familiar set-ups. They correspond, however, to the fundamental 'statistics' – as they are called by physicists – of Maxwell-Boltzmann (case (*a*)), Fermi–Dirac (case (*b*)) and Bose–Einstein (case (*c*)).

For all these cases, we can think in terms of an urn containing

$$g = g_1 + g_2 + \ldots + g_r$$

balls of r different colours, and then, with respect to different procedures for drawing a total of n balls, we seek the probabilities that the numbers of balls drawn of each of the different colours will be n_1, n_2,..., n_r. One should bear in mind, however, that there are many other interpretations that could be considered: for example, how many objects, out of a total of n, will be attributed to individuals (or placed into cells) identified by *colours* 1, 2,..., r (i.e. *associated with* balls of these colours). In practice, the individuals could be characterized in any way whatsoever: nationality, sex, marital status, school and so on (in the case of cells, it might be energy levels). If we stick to colours, this has the advantage of making it clear that, so far as the considerations we are interested in are concerned, the nature of the characteristic on which the classification is based is irrelevant (whereas, of course, this is no longer the case if one wishes to study the particular aspects of some given application).

10.3.3. We now consider the three cases mentioned above. They differ in the form of procedure used in drawing the balls; these correspond to (*a*) with replacement, (*b*) without replacement, (*c*) double replacement, terms which will be made more precise as we go along (equal probabilities being assumed throughout).

(*a*) *With replacement.* We perform n drawings from an urn with replacement. Thinking in terms of our alternative interpretation, we draw n objects in succession, distributing them among the g individuals (or cells) regardless of whether the latter have previously received any or not. This is the obvious extension to higher dimensions of the binomial distribution and is known as the *multinomial distribution*. At each drawing (independently of the previous outcomes) the probabilities of the various colours are given by $P_k = g_k/g$ (either referring to a drawing of that colour, or in favour of some

individual, or cell, identified by that colour). The probability of the various colours appearing $n_1, n_2,\ldots, n_r$ times is therefore given by

$$\omega^{(n)}_{n_1,n_2,\ldots n_r} = \frac{n!}{n_1!n_2!\ldots n_r!} p_1^{n_1} p_r^{n_2} \ldots p_r^{n_r} = \frac{n!}{g^n} \prod_k \frac{g_k^{nk}}{n_k!} = K \prod_k \frac{g_k^{n_k}}{n_k!}. \tag{10.18}$$

Special case. Taking all the $g_k = 1$ (for example, if all balls, individuals, or cells, are of a different colour; i.e. if we are dealing with the distribution among the g different balls, individuals, or cells, without speaking of colours), we obtain

$$\frac{n!}{n_1!n_2!\ldots n_r!}\left(\frac{1}{g}\right)^n = \frac{n!}{g^n}\prod_k \frac{1}{n_k!}. \tag{10.19}$$

10.3.4. (*b*) *Without replacement*. We perform n drawings from an urn without replacement. Thinking in terms of our alternative interpretation, we draw n objects in succession, distributing them only among those of the g individuals who have not yet received any. In this way, we exclude the possibility of an individual (or cell) receiving more than one object (we must therefore assume $n \leqslant g$, and we certainly have $n_k \leqslant g_k$, $k = 1, 2,\ldots, r$). This is the obvious extension to higher dimensions of the hypergeometric distribution; we obtain

$$\begin{aligned} \omega^{(n)}_{n_1,n_2,\ldots,n_r} &= \binom{g_1}{n_1}\binom{g_2}{n_2}\ldots\binom{g_r}{n_r}\Big/\binom{g}{n} = K\prod_k \binom{g_k}{n_k} \\ &= K\prod_k \frac{g_k(g_k-1)\ldots(g_k-n_k+1)}{n_k!}. \end{aligned} \tag{10.20}$$

In fact, $\binom{g}{n}$ is the number of ways in which n individuals can be chosen out of g (i.e. of distributing n objects among them, not more than one to each). The interpretation of $\binom{g_k}{n_k}$ for colour k is similar and gives the number of ways in which a distribution of the given form can take place.

Special case (as above, $g_k = 1$). The possible distributions correspond to the various possible choices of n out of the g balls (or individuals, or cells), and these number $\binom{g}{n}$. They all have the same probability, $1\Big/\binom{g}{n}$, because $\binom{g_k}{n_k}$ is either equal to $\binom{1}{1}$ or $\binom{1}{0}$, and is therefore equal to 1.

10.3.5. (*c*) *Double replacement*. We perform n drawings from an urn, replacing, on each occasion, the ball drawn, together with a further ball of the same colour (so that, after $m = m_1 + m_2 + \ldots + m_r$ drawings of the balls of various colours, the urn contains $g + m$ balls, of which $g_k + m_k$ are of colour k). Thinking in terms of our alternative interpretation, we could imagine that every individual participates at each drawing as though it were a raffle and, together with his original ticket, has a number of additional tickets, one for each object received so far.[5]

5 A somewhat more expressive example is the following. The r original individuals act as recruiting officers for companies. New individuals are assigned to companies by randomly selecting someone already present, and then assigning the individual to his company (so that, at any given moment, the largest company has the highest probability of recruiting).

In this case, we have

$$\omega^{(n)}_{n_1,n_2,\ldots,n_r} = \frac{n!}{n_1!n_2!\ldots n_r!}\frac{\prod k\, g_k\,(g_k+1)(g_k+2)\ldots(g_k+n_k-1)}{g(g+1)(g+2)\ldots(g+n-1)} = \frac{1}{\binom{g+n-1}{n}}\prod_k \binom{g_k+n_k-1}{n_k}. \tag{10.21}$$

To see this, note that the ratio giving the second factor is precisely the probability of obtaining the required distribution in some preassigned order. In fact, if we write, for example,

$$\frac{g_1}{g}\cdot\frac{g_3}{g+1}\cdot\frac{g_3+1}{g+2}\cdot\frac{g_2}{g+3}\cdot\frac{g_2+1}{g+4}\cdot\frac{g_3+2}{g+5}\cdot\frac{g_1+1}{g+6}\cdot\frac{g_2+2}{g+7}\cdot\frac{g_2+3}{g+8}, \tag{10.22}$$

we are expressing, as a product (compound probability), the probability of obtaining, in $n = 9$ drawings, colour 1 twice, colour 2 four times, colour 3 three times, in the order 1–3–3–2–2–3–1–2–2. For a different order, we merely permute the numerator; the denominator does not change. If the order is not taken into account, the required probability is that given above multiplied by the number of permutations (in which the order is preserved *among* g_k, $g_k + 1$ etc.). In the example, the number of permutations is 9!/2!4!3!; in the general case, we have $n!/n_1!\ldots n_r!$, as in equation 10.21.

Pólya's urn scheme (*for* '*contagious diseases*'). The process that we have just considered – drawings with double replacement – is known as Pólya's urn scheme (especially in the case $r = 2$, black and white balls), having been introduced by Pólya as a particular model for the spread of 'contagious diseases' (in the sense that the more a colour turns up, the more probable it is to do so again). We observe that, contrary to what one might think initially, results that differ only in the order (permutations!) have the same probability (as we saw in the case of equation 10.22). On the other hand, this also holds in the case of drawings without replacement and in other variants: for example, after each drawing replacing c balls of the colour just drawn and d balls of the other colour. If $d > 0$, we have the possibility of dealing with other cases besides the 'contagious' form. If negative values are also permitted for c and d, many conclusions still hold, but the process may – and sometimes certainly does – terminate after a finite number of drawings (it suffices to consider the case $c = -1$ and $d = 0$; the case of drawings without replacement). We could also generalize beyond the model of balls in an urn and take c and d as noninteger parameters for determining the successive probabilities.

Special case (as above, $g_k = 1$). In this case, we have $g_k + n_k - 1 = n_k$, hence the product in equation 10.21 is equal to 1 (all factors are of the form $\binom{n_k}{n_k} = 1$) and all possible distributions (with any given g and n) have the same probability:

$$1 / \binom{g+n-1}{n}. \tag{10.23}$$

We recall that $\binom{g+n-1}{n}$ is the number of ways of distributing g objects among n individuals, two distributions being considered distinct only if they differ in the *number* of objects (and not in the *particular* objects) attributed to each individual.[6]

6 See the Remark in Section 10.4.1.

Sometimes, one refers to 'the different distributions that arise when the objects are considered as indistinguishable'; in the interpretation of cases in physics where this turns out to be applicable (experimentally) it is attributed to the fact that the particles in question are 'indistinguishable'. The same interpretation also holds in the general case (g_k arbitrary), and the explanation is practically identical.

However, the interpretation in terms of the Bayes–Laplace scheme (which we shall meet in Chapter 11, 11.4.3) is possibly more satisfactory and might also be considered.

10.3.6. *Remark*. In the case of the applications in physics to which we have referred, case (*b*) (drawings without replacement) holds when Pauli's exclusion principle applies; it corresponds to the so-called Fermi–Dirac 'statistics', applicable to electrons, protons and neutrons (i.e. particles with semi-integer spins). Case (*c*) (double replacement) holds in all other cases and corresponds to the so-called Bose–Einstein 'statistics', applicable to photons, mesons and so on (i.e. particles with integer spin).

Case (*a*) (drawings with replacement, the Bernoulli scheme) corresponds to classical statistical mechanics (Maxwell–Boltzmann 'statistics'). According to modern theoretical physics this never applies, but it provides, asymptotically, an approximation to both (*b*) and (*c*) when the g_k are much larger than the corresponding n_k.

It would be very worthwhile to proceed further with the actual application of these ideas, principally to the problem of determining statistical equilibrium. However, this would take us well beyond our purpose in providing this introductory outline.

10.4 Some Particular Distributions: The Continuous Case

We now turn to the continuous case, where there are a number of interesting problems. We shall only be able to sample a few of them, choosing those that are best suited to illustrating certain useful techniques, and to presenting, in a simple fashion, those distributions most frequently encountered in practice.

10.4.1. *Subdivisions of an interval*. This is a continuous analogue of the problem we have just discussed in the discrete case. Instead of considering the subdivision of some given n objects into r groups, we consider the subdivision of an interval (for convenience, assumed to be of unit length) into r parts. In this way (or as a result of subdividing some other quantity), we end up with a collection of r random quantities $X_1, X_2, \ldots, X_r$, whose sum is equal to one.

There are various ways of performing such a subdivision. Of these, we shall consider one of the most straightforward and 'symmetric', and we shall give it its customary title, referring to it as 'random subdivision'. This has a certain convenience, so long as one does not attempt to read too much into the terminology, thinking of it as endowing this particular method of subdivision with some special significance, rather than being just a matter of convention.

More precisely, when we talk of *random subdivision* of an interval we mean that $r-1$ division points are chosen independently, each with a uniform distribution. Equivalently, we could say that, after having performed the subdivision into k parts, the kth division point is chosen by first choosing a subinterval – with probability of choice proportional to length – and then choosing a point within this subinterval by means of a uniform distribution over it. This formulation is a little 'artificial' if we are

considering subdivision of an interval but it still makes sense and has the advantage of also being applicable to the subdivision of an arbitrary quantity (mass, area, sum of money, amount of energy etc.). The distribution itself has constant density over the range of possible values; that is over the $(r-1)$-dimensional simplex defined by $x_k \geqslant 0$ $(k = 1, 2,\ldots, r)$ and $x_1 + x_2 + \ldots + = 1$. If $r = 3$, for example, it is uniform over the equilateral triangle, as shown in Figure 10.1.

Remark. It is instructive to point out that we are here dealing with the limit-case (as we pass from the discrete to the continuous) of the Bose–Einstein 'statistic', as considered above. In that case, in fact, the distributions of the n 'indistinguishable' objects over the g cells correspond to the $\binom{g+n-1}{n}$ ways in which n points (representing the objects) and $g-1$ division bars, together with a bar at either end (which represent the division into cells), can be arranged. For example, the distribution which results in 0, 2, 0, 3, 1 objects in the 1st, 2nd,..., 5th cells, respectively, would be represented by //**//***/*/. If the number of points n is large in comparison with the number of cells, the bars subdivide the interval in a manner very close to that described above. If we consider the distribution, it is practically uniform over the simplex because the possible points are uniformly distributed over it – the x_k are all multiples of $1/n$ – and all have the same probability (i.e. $1/N$, where $N = \binom{g+n-1}{n}$ is the total number of points). Note that the two cases are also analogous notationally (in the 'special case' we have $g = r$; the comparison with the general case led us, however, to prefer to write g rather than r).

10.4.2. *Problems relating to random subdivision* arise quite naturally and frequently in a number of applications. In order to be able to picture the distribution, it will often be useful to consider special cases where r has a small value and the simplex reduces to an interval $(r = 2)$, an equilateral triangle $(r = 3)$ or a regular tetrahedron $(r = 4)$. We find – for the same reasons, although the purpose is different – that the diagrams we require are the same as those already encountered in Chapter 5 (especially Figure 5.3b) and which represented probabilities p_h with sum equal to one.

In our case, it is the sum of the subintervals x_h that is equal to one. For $r = 3$, subdivision into three corresponds to some point in the triangle $A_1A_2A_3$ (having barycentric coordinates, x_1, x_2 and x_3, with $x_1 + x_2 + x_3 = 1$, where the x_h are the distances from the three sides and the height of the triangle is taken to be unity). Two simple examples will suffice to illustrate this form of representation and to show how, with its aid, one can obtain immediately certain conclusions which would involve heavy calculations if arrived at analytically.

In Figure 10.1a, the areas corresponding to $X_1 \leqslant x_1$, $X_2 \leqslant x_2$, $X_3 \leqslant x_3$ (for given x_1, x_2, x_3) are indicated with different forms of shading. The unshaded triangle which remains (with sides $1-x$, where $x = x_1 + x_2 + x_3$; we clearly have $x \leqslant 1$, so this does exist) represents the subdivisions in which $X_1 \geqslant x_1$, $X_2 \geqslant x_2$, $X_3 \geqslant x_3$. The probability of a subdivision for which this holds is therefore given by $(1-x)^2$ (the ratio of the area of the smaller triangle to the larger) and, by virtue of the homogeneity, one can see immediately that, for arbitrary r, the probability is equal to $(1-x)^{r-1}$.

Figure 10.1b $(X_1 > X_2 > X_3)$ illustrates the following problem. Suppose that $Z_1, Z_2,\ldots, Z_{r-1}$ are the abscissae of the $r-1$ division points arranged *in increasing order*: What are their probability distributions? We recall that the points are chosen independently and with a uniform distribution over the given interval, but that if we consider them as ordered

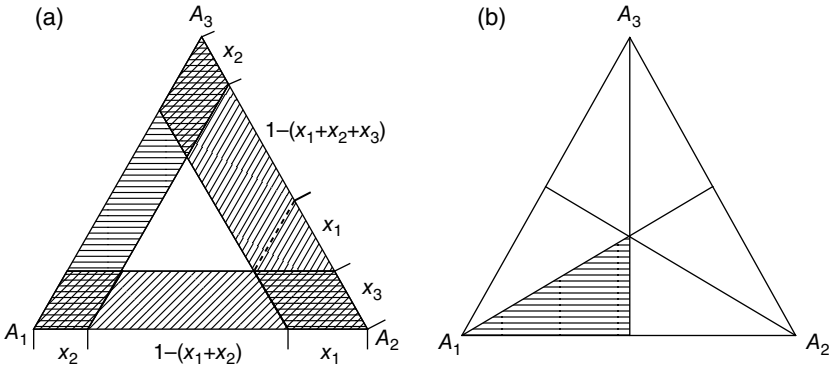


Figure 10.1 (a) The probability that $X_1 \geqslant x_1, X_2 \geqslant x_2, X_3 \geqslant x_3$. (b) The probability that $X_1 > X_2 > X_3$.

neither independence nor uniformity continues to hold. It is obvious – but one does not always think of it – that everything changes when the state of information changes (and the latter change may be obscured by the *terminology* used). As an example of this, we note that 'the 1st division point obtained' (in chronological order) may very well be 'the 7th division point obtained' (when they are taken in increasing order – for example, at some given moment when 20 of them have been considered) and that knowing these two facts to coincide changes the state of information, and with it the probability distribution. More concretely, the probability distribution of the chronologically first division point changes after each new division point is obtained if we are informed as to whether the latter is to the right or to the left of the former.

Let us first determine the distribution of the *maximum*, Z_{r-1}. To say that $Z_{r-1} \leqslant x$, amounts to saying that all the $r-1$ division points are $\leqslant x$; this has probability x^{r-1} and the distribution we are after is therefore given by

$$F(x) = x^{r-1}, \quad f(x) = (r-1)x^{r-2} \quad (0 \leqslant x \leqslant 1). \tag{10.24}$$

For the *minimum*, Z_1, we have, by symmetry,

$$F(x) = 1-(1-x)^{r-1}, \quad f(x) = (r-1)(1-x)^{r-2}. \tag{10.25}$$

In general, for the kth point, Z_k (taken in increasing order), the density is given by

$$f(x) = (r-1)\binom{r-2}{k-1}, x^{k-1}(1-x)^{r-k-1} \quad (0 \leqslant x \leqslant 1). \tag{10.26}$$

To see this, note that the probability of one (no matter which) of the $r-1$ points falling in the interval from x to $x+\mathrm{d}x$ is $(r-1)\mathrm{d}x$, which must then be multiplied by the probability that, of the remaining $r-2$ points, $k-1$ fall on the left (with probability x) and $r-k-1$ on the right (probability $1-x$).

10.4.3. *The beta distribution*. The distributions that we have just encountered belong, in fact, to the family of beta distributions, a family which finds frequent and important applications. The general form of the density is given by

$$f(x) = Kx^{\alpha-1}(1-x)^{\beta-1}\left(K = \Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta); \Gamma(n) = (n-1!)\right), \tag{10.27}$$

where α and β are positive *real numbers* (and not necessarily integers as in the previous example). If α (and/or β) is <1, the density tends to infinity at $x=0$ (and/or at $x=1$). We have already seen an example of this; $\alpha=\beta=\frac{1}{2}$ corresponds, in fact, to the arc sine distribution. In the more usual case (α and β greater than 1), the density has a maximum at $(\alpha-1)/(\alpha+\beta-2)$, and on either side of this the curve slopes downwards, reaching zero at $x=0$ and $x=1$. The prevision and standard deviation are given by $\alpha/(\alpha+\beta)$ and $\sqrt{[\alpha\beta/(\alpha+\beta+1)]}/(\alpha+\beta)$, respectively, whatever the values of α and β. Note that for a given prevision (i.e. α/β fixed), the standard deviation behaves like $1/\sqrt{(\alpha+\beta+1)}$, decreasing as α and β increase. The distribution, therefore, thickens around the prevision (and also around the mode, which differs little from the prevision and tends to it asymptotically).

10.4.4. *Extension*. The argument that we gave in the case of the division points extends immediately to the case of any n independent random quantities having the same distribution F, where F can be any distribution at all.[7]

The distribution of the maximum of $X_1, X_2,\ldots, X_n$ is given by

$$F_{(n)}(x)=F^n(x),\quad f_{(n)}(x)=nF^{n-1}(x)f(x), \tag{10.28}$$

and the density for the kth largest by

$$f_{(k)}(x)=n\binom{n-1}{k-1}F^{k-1}(x)\big(1-F(x)\big)^{n-k}f(x). \tag{10.29}$$

Similar expressions can be obtained under more general conditions.

If, for the sake of simplicity, we restrict ourselves to the maximum, we obtain, in the general case,

$$F_{(n)}(x)=F\big(x,x,\ldots,x\big), \tag{10.30}$$

where F is the joint distribution function of the n random quantities. If, in r particular, the random quantities are independent (but each X_k has a different distribution $F_k(x)$), we have

$$F_{(n)}(x)=F_1(x)F_2(x)\ldots F_n(x). \tag{10.31}$$

10.4.5. *'Random' subdivision and the Poisson process*. Suppose that, in a Poisson process, n occurrences are known, or are assumed, to have taken place in some given interval; then, in the sense we have defined, they form a 'random subdivision' of the interval. Conversely, if we imagine the subdivision of an interval of length $n+1$ by means of n points in such a way that each of the $n+1$ subintervals has expected length 1, then, as n increases, we approach a Poisson process (with intensity $\mu=1$). The distributions relating to the 1st, 2nd,..., kth, ... positions now belong to the gamma family instead of the beta as above.

In both these cases, the length of each interval is, in prevision, equal to 1. In general, however, it is important that the method of picking such an interval should be made explicit. If we refer to the 'third interval starting from 0', or 'the first interval after $x=x_0$',

7 Note that n corresponds to the 'number of division points' of the preceding case, where it was denoted by $r-1$ because there were r subintervals.

then what we have said is true. It is clearly no longer true if we look at 'the shortest', or 'the longest' (in which case, we reduce to the problem considered above, independence holding in the Poisson case, but not for a random subdivision). It is perhaps not quite so obvious that the result no longer holds if we pick out the interval 'containing some given point', but it is clear, on reflection, that this method does favour the longer intervals. In actual fact, the prevision of the length of an interval chosen in this way is *twice* that of the Poisson case, and only a little less than twice that of the case of a random subdivision. In the first case, the prevision of the distance from a given point (division point or not) to the first division point, both on the left and on the right, is equal to 1. In (the second case, the point chosen as a reference point plays the rôle of an additional 'point chosen at random'.[8] This means that the original interval, of length $n+1$, turns out to be subdivided into $n+2$ subintervals, each of which has expected length $(n+1)/(n+2)$. Two of these subintervals join together to form the interval into which the new division point has fallen, and the expected length of this interval is therefore $2(n+1)/(n+2) = 2 - 2/(n+2)$.

10.5 The Case of Spherical Symmetry

10.5.1. *Examples with spherical symmetry.* We shall obtain further useful insights by considering – in the plane, in ordinary space, and in an arbitrary number of dimensions – distributions possessing spherical symmetry. In particular, we shall consider the normal distribution. Referring to the three-dimensional case for convenience, this means that the density (provided it exists) is a function of the distance ρ only; that is $f(x, y, z) = g(\rho)$, a function of $\rho^2 = x^2 + y^2 + z^2$.

10.5.2. *Distance from the origin.* The distance $\left(X_1^2 + X_2^2 + \ldots + X_r^2\right)^{\frac{1}{2}}$ has a probability distribution with density $f(p) = Kg(\rho)\rho^{r-1}$. Taking the particular case of a uniform distribution inside the hypersphere (with radius 1), we obtain $f(\rho) = K\rho^{r-1}$ and note that this is identical to what we obtained for the distribution of the abscissa of the maximum when r points were chosen at random in [0, 1]. Observe that, for large r, the volume is concentrated near the surface; that is, for any given $\varepsilon > 0$, the layer between $1 - \varepsilon$ and 1 includes all the volume apart from a fraction θ, which tends to zero as r increases. More precisely, the distance from the surface, as r increases, tends asymptotically to an exponential distribution with prevision $1/r$.

In the case of the normal distribution, the distance is distributed with density

$$f(\rho) = K\rho^{r-1}\mathrm{e}^{-\rho^2/2}. \tag{10.32}$$

8 For this to hold exactly, we require the point to be 'chosen at random', and its position to be unknown. In other cases, the result is very little altered except when the point is very close to the end-points (and then one of the two subintervals is necessarily small). This should provide an adequate background to more complicated situations, as well as illustrating how such complications can arise in seemingly harmless formulations of problems if one is not sufficiently careful.

For $r = 3$, we note that we obtain *Maxwell's formula* for the distribution of the (absolute values of) velocities in a gas, assuming them to be normally and spherically distributed : $f(v) = Kv^2 e^{-v^2/2}$ (where we take the prevision of the square of the velocity to be equal to 3; that is equal to 1 for each component).

The distribution given by equation 10.32 is widely used in a number of problems. In particular, it occurs in statistics, where one often takes as a basis of comparison the square of some deviation from a 'true' value. In this case, it is known as the χ^2 ('*chi-square*') distribution. If we take $x = \rho^2$ as the variable rather than ρ, we obtain a gamma distribution

$$f(x) = Kx^{(r-2)/2} e^{-x/2}.$$

In fact, if we temporarily write $f_1(x)$ in order to avoid confusion with $f(\rho)$, we have

$$f_1(x)\mathrm{d}x = f(\rho)\mathrm{d}\rho^{9} = Kf(x)\mathrm{d}x = K.x^{(r-1)/2} e^{-x/2} x^{-\frac{1}{2}}\mathrm{d}x$$

(the constant ½ being included in K).

10.5.3. *Distance from a hyperplane to the origin* (or, alternatively, the coordinate, or projection, onto an arbitrary axis).[10] This has as its distribution the projection of the spatial distribution and is the same for all axes. In other words, if $f(x)$ gives the density of the distribution of X, then it also gives that of Y and Z, and of any other coordinate $aX + bY + cZ$, where $a^2 + b^2 + c^2 = 1$ (i.e. with the same unit of measurement). Given $g(\rho)$, we have

$$f(x) = K\int_0^\infty g\left(\sqrt{\left(x^2 + \lambda^2\right)}\right)\lambda^2 \mathrm{d}\lambda, \tag{10.33}$$

and, for general r,

$$f(x) = K\int_0^\infty g\left(\sqrt{\left(x^2 + \lambda^2\right)}\right)\lambda^{r-1} \mathrm{d}\lambda. \tag{10.34}$$

We have already seen that in the case of the normal distribution (and only in this case) g and f coincide (up to the normalization constant). We have also seen that for a uniform spherical distribution, $g(\rho) = K > 0$ for $\rho \leqslant 1$, $g(\rho) = 0$ for $\rho > 1$, we have $f(x) = K(1 - x^2)^{(r-1)/2}$ (see Section 7.6.8), and we observe that we are dealing with a beta distribution,

$$f(x) = K\left(1 + x\right)^{(r-1)/2}\left(1 - x\right)^{(r-1)/2},$$

defined over [–1, 1], rather than over [0, 1]. Let us take up again the case of a distribution on the surface of a unit sphere; more precisely, we shall consider a spherical layer (points whose distances from the origin lie in the range $1 - \varepsilon$ to 1) whose thickness, $\varepsilon > 0$, we let tend to zero. In this way, we obtain

---

9 We take this opportunity of pointing out how a change of variable leads to an altered form of density (obviously: we are dealing with the derivative of a function of a function!). For increasing transformations, this applies directly: for transformations which are not one-to-one, we have to add up the separate contributions. As a practical rule, it is convenient to transform (as we have done) $f(y)\,\mathrm{d}y$ into $f_1(x)\,\mathrm{d}x$, rather than writing $f_1(x) = f(y).\ \mathrm{d}y/\mathrm{d}x$.

For transformations of several variables one proceeds in a similar fashion, but multiplying by the Jacobian $\partial(y_1,\ldots, y_r)/\partial(x_1,\ldots, x_r)$ instead of by $\mathrm{d}y/\mathrm{d}x$.

10 This differs only that, in speaking of *distance*, one needs to take the *absolute value* of the abscissa. Given the symmetry, the density is $2f(x)$ for $x \geqslant 0$, and zero for $x \leqslant 0$, rather than $f(x)(-\infty < x < +\infty)$.

$$\begin{aligned} f(x) &= K\left[\left(1-x^2\right)^{(r-1)/2} - \left(\left[1-\varepsilon\right]^2 - x^2\right)^{(r-1)/2}\right] \\ &\simeq K\left[2(r-1)\varepsilon\left(1-x^2\right)^{(r-1)/2-1}\right] = K\left(1-x^2\right)^{(r-3)/2}. \end{aligned} \tag{10.35}$$

We are thus led to the same distribution, but with r reduced by 2. For the particular case $r=2$, we have $f(x)=K/\sqrt{(1-x^2)}$; in other words, as was obvious geometrically, we again obtain the *arc sine* distribution. For $r=3$, we obtain the uniform distribution (as one would expect from the well-known relation between the area of the sphere and of the cylinder). In both of these cases, as in many other cases of this kind, as r increases the projection of the distribution tends to normality.

The distance from a straight line (or plane, or arbitrary Euclidean space with dimension $d<r$) passing through the origin can be shown to lead to a gamma distribution (with parameters $\alpha=d$ and $\beta=r-d$).[11]

10.5.4. Finally, let us consider the *central projection* of a distribution with spherical symmetry (r-dimensional) onto an arbitrary hyperplane ($(r-1)$-dimensional); a straight line if $r=2$, a plane if $r=3$ and so on. This is clearly the same no matter which hyperplane we take (apart from changes of scale, which can be avoided in any case if we adopt the convention of taking the hyperplane to be unit distance from the origin) and *no matter what distribution one starts with* (in other words, it does not matter what $g(\rho)$ is: in fact, it does not matter how the mass moves along the radii of the projection). We might as well assume, therefore, that the mass is uniformly distributed on the surface of a hypersphere with radius 1 (centred at the origin).

We have just seen, however, that the projection of this distribution onto an axis, $x=\cos\phi$ (Figure 10.2), has density $K(1-x^2)^{(r-3)/2}$. A mere change of variable suffices, therefore, to obtain the distribution in terms of either the angle ϕ, or $y=\tan\phi$, or $z=1/y=\cot\phi$.[12]

From $x=\cos\phi$, we obtain

$$\left(1-x^2\right)^{\frac{1}{2}} = \sin\phi,\ \mathrm{d}x = \sin\phi\,\mathrm{d}\phi,$$
$$K\left(1-x^2\right)^{(r-3)/2}\mathrm{d}x = K\sin^{r-3}\phi.\sin\phi\,\mathrm{d}\phi = K\sin^{r-2}\phi\,\mathrm{d}\phi,$$

the distribution for ϕ having density proportional to $\sin^{r-2}\phi$ (i.e., as is well known from geometry, the area of the ring cut on the hypersphere by cones with semi-angles ϕ and $\phi+\mathrm{d}\phi$).

From $y=\tan\phi$, that is $\phi=\tan^{-1}y$, we obtain

$$\sin\phi = y\left(1-y^2\right)^{-\frac{1}{2}}, \mathrm{d}\phi = \left(1-y^2\right)^{-1}\mathrm{d}y,$$
$$K\sin^{r-2}\phi\,\mathrm{d}\phi = Ky^{r-2}\left(1+y^2\right)^{-r/2}\mathrm{d}y.$$

11 This problem crops up in connection with problems in theoretical physics; see J. von Neumann, *Zeitschr. Phys.*, **57** (1929); A. Loinger, *Rend. S.I.F.*, 1961.

12 The letters y and z are used here simply for convenience and not in their usual sense of coordinates.

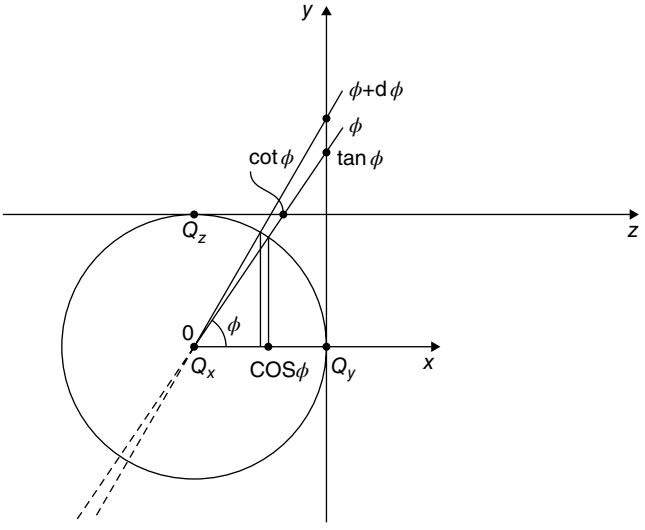


Figure 10.2 The central projection (origin 0) of a spherically symmetric distribution. The functions ϕ, cos ϕ, tan ϕ, cot ϕ, and their derivatives, appear in the various problems which we consider.

Finally, from $z = 1/y$, that is $y = z^{-1}$ we obtain

$$\begin{aligned} \mathrm{d}y &= -z^{-2}\mathrm{d}z, \\ Ky^{r-2}\left(1-y^2\right)^{-r/2}\mathrm{d}y &= Kz^{-(r-2)}\left(1+z^{-2}\right)^{-r/2}z^{-2}\mathrm{d}z \\ &= Kz^{-r}\left(1+z^{-2}\right)^{-r/2}\mathrm{d}z = K\left(1+z^2\right)^{-r/2}\mathrm{d}z. \end{aligned}$$

If X_1, X_2,..., X_r are random quantities whose joint distribution has spherical symmetry, and we set $X = X_1$, $R =$ distance of the point $(X_1,\ldots, X_r)$ from $0\left(R^2 = X_1^2 + X_2^2 + \ldots + X_r^2\right), D = \sqrt{\left(R^2 - X^2\right)} =$ distance of the same point from the x-axis, then the variables previously denoted by x, y, z and ϕ correspond to X/R, D/X, X/D and $\tan^{-1}(D/X)$, respectively. Their distributions, therefore, have densities of the form:

$$X/R \qquad (\cos\phi): f(x) = K\left(1-x^2\right)^{(r-3)/2}\left(-1 \le x \le 1\right) \tag{10.36}$$

$$D/X \qquad (\tan\phi): f(x) = Kx^{r-2}\left(1-x^2\right)^{-r/2} \tag{10.37}$$

$$X/D \qquad (\cos\phi): f(x) = K\left(1+x^2\right)^{-r/2} \tag{10.38}$$

$$\tan^{-1}(D/X) \quad (\phi): f(x) = K\sin^{r-2}x\left(-\pi/2 \leqslant x \leqslant \pi/2\right); \tag{10.39}$$

where x, as usual, denotes the variable. We note that in the case of D/X with r odd, we would have to include the absolute value sign, or, alternatively, think of K changing its sign as x does. In all cases, the same distributions (if we double K and restrict the range to $x \geqslant 0$) correspond to the absolute values of the random quantities ($|X|/R$ etc.).

The distribution of $D/|X|$ is that of the distance for the distribution projected onto the hyperplane. Dividing by $x^{(r-1)-1}$, we obtain the $(r-1)$-dimensional density as a function of x, corresponding here to the distance ρ. We can, therefore, write $g(\rho) = K(1+\rho^2)^{-r/2}$; formally, this is the same expression as that which is given for $|X|/D$, but the meaning is different, and K appears because we are dealing with an $(r-1)$-dimensional distribution instead of the one-dimensional case. Note that the exponent should be $-(r+1)/2$, corresponding to r being the dimension of the space we are dealing with, rather than that from which we have projected. In the simplest case of the projection of a plane distribution with circular symmetry onto a straight line ($r=2$, distribution of Y/X; the reader should be able to deduce this result directly from an inspection of the diagram), we obtain the *Cauchy distribution*:

$$f(x) = K/\left(1+x^2\right) \ \left(K = 1/\pi\right), \ F(x) = \frac{1}{2} + (1/\pi)\tan^{-1} x. \tag{10.40}$$

This is the most direct characterization of the Cauchy distribution (which is usually presented as the special case of Y/X with X and Y independent, centred normals, $m=0$). As we have seen already, this is a stable distribution with infinite variance.

Notice that for D/X we have infinite variance for every r, whereas for X/D we have $f(x) \sim x^{-r}$ and hence no moments of order $\geqslant r-1$ (they become infinite). This latter distribution (X/D, with r arbitrary) is also of great importance in statistics, where it finds wide application as *Student's distribution* (Student being the nom-de-plume of W.S. Gosset, who introduced it into statistics; see Chapter 12, 12.3.6).

11

Inductive Reasoning; Statistical Inference

11.1 Introduction

11.1.1. Within the ambit of the logic of certainty, that is to say ordinary logic, valid arguments are deductive arguments. Conclusions which are *certain* can only be arrived at by establishing that they are implicit in something already known. In other words, we arrive at the particular through the general. In doing so, however, it is clear that we can never enlarge our field of knowledge (except in the sense that certain features of our previously acquired knowledge, of which, perhaps, we were previously unaware, are now made more explicit).

The form of argument leading to conclusions that go beyond what is already known, or what has previously been ascertained, is different; this is the so-called inductive form of argument. We have used 'so-called' because, in fact, we must first of all discuss whether, and in what sense, it is legitimate to refer to it as a form of 'argument' at all (see Sections 11.1.3–11.1.4).

The problem of induction arises in every field and at every level: from the examination of arguments for and against various scientific theories, to those concerning the guilt of someone suspected of a crime; from methods for establishing, on the basis of some given data, the conditions for a specific kind of insurance policy, to methods of estimating some quality or other to the required degree of accuracy on the basis of measurements which are inherently imprecise.

It is particularly instructive to consider the process by which new scientific theories are formulated. The first step is an intuitive one, arising out of some particular set of observations, but then various modifications are made as a result of more up-to-date results, which suggest that this or that alternative theory provides a better explanation. In essence, it is always a question of analysing the current state of information by means of Bayes's theorem (except that, in this rather open-ended and imprecise context, such applications of the theorem are necessarily qualitative in nature). The most interesting feature of all this is, perhaps, the substantial scope which is left for the personal judgements of individual scientists. In particular, it is interesting to note the prevalent, conservative aversion to any form of novelty; an aversion which some might regard as an alarming symptom of the superstitious faith placed in the 'scientific truths' of the moment.

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

There are two common fallacies which deserve special mention. One consists in believing that a theory can be disproved merely by discrediting some particular explanation, consequence, or application of it. This is not so: it may well be the case that the particular explanation is not essential, or that the particular application breaks down for some other reason.[1] The cursory manner in which new ideas are discussed is to be deplored, because such ideas, even though they may turn out to be false trails, usually contain within them the germ of something fruitful. In this respect, the second of the two fallacies is even more dangerous. This fallacy consists of leaving out of consideration certain of the data or observations. In the logic of certainty this is quite legitimate: it is perfectly proper to start from some restricted set of hypotheses and to *deduce* the corresponding restricted set of conclusions (which are, in any case, *correct*). In the logic of probability this is not so (as is obvious – even without a consideration of *likelihood*, Chapter 4, 4.6.1 – if one considers the bias that would be introduced if all the evidence against some particular hypothesis were suppressed). It is important to note how easy it is to overlook this fact – albeit inadvertently!

A deeper analysis of the way in which scientific thought evolves, in all its many aspects, would make an extremely interesting study, although one fraught with difficulty. In actual fact, I do not know of any work along these precise lines, nor of any such attempt. It would need to be the history of a continuous series of conceptual U-turns, occasioned by singular minds reflecting upon singular results, and initially greeted with hostility, incomprehension and suspicion, until, finally, the weight of favourable evidence and the resulting improvement in the theoretical formulations renders them acceptable.

We can quote a few examples of this. They may serve to give some idea of how a few of the main aspects of the problem should be tackled in the context of a synthesis that captures the essence of the whole. In the work of Weisskopf (which we have already quoted in the footnote to Chapter 8, 8.8.4), the decisive part played in every field by revolutionary conceptual innovations and changes is stressed and allotted its rightful place within an overall examination of the development of modern scientific conceptions. The intuitive basis of an idea and the way in which it develops as one searches for evidence supporting it is vividly described by James D. Watson in *The Double Helix*, Weidenfeld and Nicholson, London (1968), an autobiographical description of the events leading up to the discovery of the structure of DNA (the substance of which the genetic code and so on is made up). A critical analysis of the academic establishment's attitude to 'disturbing' theories can be found in the article The scientific reception system' by Alfred de Grazia, in the volume *The Velikovsky Affair: The Warfare of Science and Scientism,* University Books, New York (1966), which he edited. Various considerations closely bound up with the themes of this book are developed in a paper of mine, 'Remore e freni sul cammino della scienza', appearing in *Civiltà delle macchine* (1964).

---

1 Here are two examples. Wegener's theory (of 'continental drift') has been rejected on the grounds that the mechanism he suggested by way of explanation is not appropriate. But this in no way excludes the possibility of the theory being correct (with the explanatory detail revised, under the assumption of some other mechanism). Velikovsky's theory (concerning certain aspects of the planetary system) was considered absurd because, among other things, it implied that the temperature of Venus could be ridiculously high.... Such temperatures were, in fact, confirmed by Mariner II, and by other observations. This in no way proves the correctness of the theory (which is rather speculative – at least, in terms of current views) but it is sufficient to discredit the claims of those scientists who believe they have the right to make their superficial judgements, without even bothering to examine the numerous, careful arguments put forward.

Everyone is familiar with the background to the struggle for the establishment of new ideas; from relativity theory to quantum physics, from the theory of evolution to that of Mendelian heredity. It would be instructive to obtain a critical compilation of all this material in order to ascertain whether, and to what extent, the situation (mutatis mutandis) has improved since the time of Galileo.

Another aspect of the problem, and one more strictly in line with the subject matter of this work, and, in particular, of this present chapter, is that of examining these processes of discovery and acceptance in the context of the probabilistic basis of the inductive argument. As we have remarked already, we are dealing with rather imprecise situations, so that any estimation of the probabilities involved could only be attempted by experts attempting to put themselves in the place of the scientists of the period in question, assuming only the knowledge available to them. It might be possible to do something worthwhile in this connection.[2] There is a vast literature in this area but, in my opinion, it is inspired by considerations that are too abstract and formalistic (in particular, this applies to the work of R. Carnap and K. Popper).[3] In contrast, the critical comments of H. Jeffreys in 'Logic and scientific inference' (which forms Chapter I of his book *Scientific Inference*, cited in the footnotes to Chapter 7, 7.5.5) are beautifully penetrating and witty. By means of a brilliant, imaginary dialogue between a logician and a botanist, he attempts to establish the inevitably uncertain and tentative nature of all scientific 'truths' or 'laws' and, for this reason, the necessity of making probabilistic logic the basis of every argument.[4] However, the treatment stops short of actually using this idea, or examining it more deeply; neither does it mention the possibility of applying it to the grander problem of synthesis; that is, to the problem of choice among competing theories.

11.1.2. The range of problems for which the inductive argument can be carried out specifically as an application of the calculus of probability in a technical sense is more modest and concerns rather special problems arising in the context of an already accepted approach. We shall confine ourselves to these kinds of problems and, in particular, to the most basic and straightforward cases. We begin by identifying, rather crudely, three possible meanings, or forms, of induction; the second form is the one where we encounter the standard applications, and we shall concentrate on this one.

- *First form.* Here we obtain conclusions of a (more or less strictly) *deterministic* kind. From the realization that in some given number of more or less similar cases some given event has always occurred in precisely the same way, we are often led to expect that the same thing will continue to happen in the future, or even to believe that it must necessarily happen on account of there being some 'law'. This is the most extreme case.

---

2 Research carried out by K. Pearson and G.M. Morant, in which they classify and evaluate all the ingredients which could be put forward in any discussion concerning the authenticity of Cromwell's skull (and the conclusion reached is that, to all intents and purposes, it is authentic), might, in some sense, serve as an example which could possibly be extended to more interesting problems. See quotations and comments in B. de Finetti and L.J. Savage, 'Sul mondo di scegliere le probabilità iniziali', in *Bibl. del Metron*, C, Vol. I, Rome (1962), pp. 130–131 and 153.

3 A comparative and critical study of the various ideas put forward in this area can be found in a useful paper by Imre Lakatos; 'Changes in the problem of inductive logic', in *The Problem of Inductive Logic*, North-Holland Publ. Co., Amsterdam (1968).

4 See similar considerations put forward in the Appendix (especially Section 13).

- *Second form.* From the realization that some given event has occurred in some given way *almost always,* or *with some given frequency* (e.g. 37·2%), we are often led to expect that in the future it will continue to happen almost always, or with that given frequency, as the case may be. This is the most typical example of *statistical inference* as it is commonly understood.
- *Third form.* From the knowledge of the behaviour of some given event in a collection, or sequence, of more or less similar cases in the past, we are often led to make some kind of forecast of the future. For example: that an apparent tendency for a frequency to decrease will continue; or that this will apply to the tendency of successes to group together in runs, or to alternate with failures; and so on. This can be regarded as the most general case.

We cannot claim that there is any really clear-cut distinction among the three cases (especially between the last two), nor that the distinction between past and future has any real importance. The inductive argument can equally well be used to make conjectures about behaviour at or before the time for which observations are available. The three categories should, therefore, simply be treated as a convenient way of concentrating attention on certain aspects of the problem and for reflecting upon the questions raised.

If one were interested in the difficult and complex questions raised by considering the notion of indeterminism (*marginal* or *static* – with or without experimentation – or *dynamic,* using stochastic models), one might adopt a different classification (for example, that suggested by J. Neyman).[5] We shall not deal with these problems, however.

11.1.3. If by 'argument' we mean something based upon logic – the logic of certainty, ordinary logic – then it is clear that the 'inductive argument' is not an 'argument'. Even in the first form of induction, where the conclusion (whether valid or not) has the logical meaning of a statement, it is clear that logic cannot provide any proof of its validity. Knowing that an event has never occurred in the past in no way excludes the possibility of its occurring in the future (even if we admit, in order to pre-empt any hair-splitting objections, that an event which has never been observed has, in fact, never occurred). So far as the other forms are concerned, there is always the objection that from the knowledge of the past (or, at least, of that which has already been observed or ascertained) nothing can be logically concluded concerning the future (or, in general, concerning that which is as yet unobserved or unknown and which could be anything at all, even the unimaginable). Moreover, in these cases, the 'conclusions' themselves do not even have the logical status of precise statements (leaving aside the question of their validity).[6]

Within what 'logical' ambit, then, might it be admissible to assert that the 'inductive argument' is an 'argument'? From our standpoint the answer is straightforward. It is

---

5 Jerzy Neyman, 'Indeterminism in science', etc., in *Jour. Amer. Math. Ass.* (1960); see comments in B. de Finetti and F. Emanuelli, *Economia delle assicurazioni*, Vol. XVI of *Tratto italiano di economia* (edited by C. Arena and G. Del Vecchio), Utet, Torino (1967), pp. 17–19.

6 In the 'third form', we could also encounter statements of a deterministic kind, expressing, for example, a tendency to decrease or fluctuate according to some precisely stated law (like the exponential, or the sine curve, etc.) suggested by extrapolation. In this case, the second objection (that of imprecision) no longer holds; a third objection arises, however (or one might say that the first objection becomes more serious), because of the large degree of arbitrariness that attaches, in general, to the choice of an extrapolation formula.

admissible within the ambit of probabilistic logic; that is, that of (subjective) probability theory, which, for us, is the only form of logic required over and above that of ordinary logic. In fact, in what follows (in this and the final chapter) we shall illustrate all the questions that arise in the context of induction by presenting them within the framework of the subjectivistic–probabilistic interpretation. The vital element in the inductive process, and the key to every constructive activity of the human mind, is then seen to be Bayes's theorem.

11.1.4. Those who do not accept this point of view (and they are, unfortunately, in the majority) come up against a dead-end and are in a different situation. If one accepts, in its totality, the subjectivistic interpretation, probability theory constitutes the logic of uncertainty; this complements the logic of certainty and the two together form a unified and complete framework within which to conduct any argument. Those who reject this point of view find themselves without any coherent foundation on which to build. Between the logic of certainty and probability theory – reduced now to a fragmentary collection of those aspects that can be provided with an objectivistic disguise – there is a void; any attempt to fill this must be without foundation and consists, in the final analysis, of empty phrases. A useless attempt is made to enlarge and extend the rôle of the calculus of probability (and the applications thereof – referred to nowadays by the Anglo-Saxon title of 'statistics') in a manner that cannot be justified within the terms of the objectivistic assumptions and which, in any case, falls far short of the required generality, a condition met only by the subjectivistic interpretation. As is evidenced by the ever increasing proliferation of *ad hoc* methods for special cases and subcases (Adhockeries!), and the disputes to which these give rise within the ranks of the supporters of objectivistic conceptions, all such efforts fall short of being either satisfactory or sufficient. The gap remains.

In order to be able to provide 'conclusions' – but without being able to state that they are *certain*, because they are undoubtedly not so, and not wanting to say that they are *probable*, because this would involve admitting subjective probability – a search is first made for words that appear to be expressing something meaningful, it is then made clear that they do not, in fact, mean what they say, and then, finally, a strenuous attempt is made to get people to believe that it is wise to act as if the words did, in fact, have some meaning (though what it is heaven only knows!).

As examples of such *words*, we are said to '*accept*' or '*reject*' an '*hypothesis*', and to give an '*estimate*' of a quantity which is not known precisely.

In order to be dealing with a *concept* rather than a mere *word*, we should require that an *estimate* be some value arising in the context of the probability distribution attributed to the unknown quantity: for example, the prevision, the median, or whatever, especially if selected in a manner appropriate to some specific decision. If probabilities and probability distributions are not mentioned, any reference to an 'estimate' is a nonsense.

Similarly, if we use 'accept' and 'reject' to mean that the probability attributed to some given hypothesis is large enough (or small enough) for us to behave, in certain respects, as if the hypothesis were true (or false), then again we would be dealing with *concepts* and not mere words.

It is convenient at this point to enter a further reservation, this time in connection with the use of the word 'hypothesis'. What we have said above only makes sense if we are referring to an 'hypothesis' for which it is possible to verify directly whether or not it is true.

If, instead, the 'hypothesis' is somewhat of an abstraction, used solely as an interpretative device, suitable only for summarizing certain features of the problem, and depending on certain given facts, the latter neither requiring it nor capable of ruling it out, then it would be illusory, or, at least, suspect (as is the case when we ask whether 'light is a wave or particle phenomenon', or whether 'a particular individual is intelligent'). Strictly speaking, one would need to replace such statements with a precise list of the verifiable, factual circumstances which one would accept as a substitute. If the 'hypothesis' expresses an opinion about probabilities (either implicitly or explicitly) then matters are even worse. As examples, we could take the following: 'this coin is perfect, in the sense that $p = \frac{1}{2}$'; 'sunspots influence economic life (in the sense of there being a probabilistic correlation)'; 'the fact that having been the cause of a car accident increases the risk of one's being involved in other accidents in the future'. In these cases, it would be necessary to substitute formulations – if any such exist – that could be expressed in terms of (subjective) probabilities referring exclusively to facts and circumstances that are directly verifiable and of a completely objective, concrete and restricted nature.

11.1.5. The final sentence above re-emphasizes something we have pointed out on numerous occasions. The objectivistic conception of probability and statistics, by misguidedly attempting to make everything objective (including things which cannot be so), in fact has the opposite effect: instead of objectivity being granted its rightful, important place, it is discredited by being claimed in contexts where it is inappropriate. The same thing would happen if someone tried to raise the status of the property of 'rigidity' by referring to all solid bodies as 'rigid bodies' (including those which are elastic or plastic). The effect would be to deprive the notion of 'rigidity' of any meaning or applicability, even in those situations for which it was originally introduced and served a useful purpose, and where it needs to be free of any distortions of meaning, ambiguities or artificial interpretations.

In the philosophical arena, the problem of induction, its meaning, use and justification, has given rise to endless controversy, which, in the absence of an appropriate probabilistic framework, has inevitably been fruitless, leaving the major issues unresolved. It seems to me that the question was correctly formulated by Hume (if I interpret him correctly – others may disagree) and the pragmatists (of whom I particularly admire the work of Giovanni Vailati[7]). However, the forces of reaction are always poised, armed with religious zeal, to defend holy obtuseness against the possibility of intelligent clarification. No sooner had Hume begun to prise apart the traditional edifice, then along came poor Kant in a desperate attempt to paper over the cracks and contain the inductive argument – like its deductive counterpart – firmly within the narrow confines of the logic of certainty.

The remainder of this work can be seen as an attempt to do away with such nonsense once and for all. In both the general philosophical context, and in the more technical

7 See G. Vailati, *Scritti* (edited by Seeber), Florence (1911). Giovanni Vailati, a mathematician of the Peano school, was an original, profound and committed supporter of pragmatism in Italy (which had several features – which I, in fact, approve of – distinguishing it from the American version of Peirce, James etc.). The beginnings of a work on pragmatism (which was to be with Mario Calderoni, but was unfinished because of Vailati's death) are published in two articles (CCX and CCXI) to be found in the above volume (pp. 420–432 and 933–941): 'Le origini e l'idea fondamentale del pragmatismo' and 'Il pragmatismo e i vari modi di non dir niente'. See, also (CLIX, pp. 684–694), 'Pragmatismo e logica matematica'.

mathematical–statistical sense, we shall try to show that these questions, which are, in themselves, perfectly clear and straightforward, can be formulated in a perfectly clear and straightforward manner. All that is required is that we abandon the traditional pursuit of creating for ourselves pretentious and misleading malformations.

11.2 The Basic Formulation and Preliminary Clarifications

11.2.1. In our formulation, the problem of induction is, in fact, no longer a problem: we have, in effect, solved it without mentioning it explicitly. Everything reduces to the notion of conditional probability (introduced in Chapter 4) and to the considerations that were developed there (particularly in Chapter 4, Section 4.14, albeit rather concisely) concerning 'stochastic dependence through an increase in information'.

What is required now is a more systematic study of this topic, oriented specifically towards the questions which presently concern us. These questions only differ from the general case – that of the 'effect of an increase in information' – insofar as the information in our case may well be obtained by design, by means of observations, or even through appropriate experimentation. However, this distinction is of no real importance.

11.2.2. By virtue of having observed, or having obtained the information, that some given complex A of events has occurred, what *are* we entitled to say about some future event E? (Or about some collection of future events? Or about events for which 'future' is replaced by 'not yet known'?) The answer is … nothing! Nothing 'certain', that is, because nothing justifies our making any *prediction* about a future event E unless it is assumed to fall within the ambit of some never-failing 'laws' (this might apply, for example, to an eclipse of the sun; even in this case, however, if one wished to be rigorous it would be necessary to add 'assuming no violent changes to the planetary system, outside of what previous observations could have led us to expect', and, as a blanket qualification, 'unless the above-mentioned laws are disproved'). And nothing can be said, in any objective sense, concerning probability or *prevision.* This means that no restrictions can be made: every prevision – that is, every evaluation of probability – can be made freely and is entirely a matter for the subjective judgement of the individual.

It is only *within* this process of subjective judgement that certain restrictions occur. These are the restrictions imposed by coherence, from which derives all that can legitimately be said concerning 'inductive reasoning', and which essentially reduces to the theorem of compound probabilities (or to its corollary, Bayes's theorem; the latter often being the more expressive form).

Suppose that $\mathbf{P}(E)$ represents the probability evaluated on the basis of our assumed information; that is that of knowing of the occurrence of the complex of events A (perhaps by means of certain observations). If H_0 denotes the entire complex of initial information (see Chapter 4 Section 4.1), and $\mathbf{P}^0(E) = \mathbf{P}(E|H_0)$ denotes the corresponding probability, then, in fact, $\mathbf{P}(E) = \mathbf{P}(E|H_0A)$, the probability corresponding to the original information plus that provided by the knowledge of A: Bayes's version of the theorem of compound probabilities implies in this case that we must have

$$\mathbf{P}(E) = \mathbf{P}^0(EA) / \mathbf{P}^0(A) = \mathbf{P}^0(E)\mathbf{P}^0(A \mid E) / \mathbf{P}^0(A). \tag{11.1}$$

Remarks. There is a delicate point here which requires some attention. When we defined conditional probability (Chapter 4, Section 4.1), we stated that the H appearing in $\mathbf{P}(E|H)$ means that this is the probability You attribute to E if '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*)'. It would be wrong, therefore, to state, or to think, in a superficial manner, without at least making sure that these explanations are implicit, that $\mathbf{P}(E|H)$ is the probability of E once H is known. In general, by the time we learn that H has occurred, we will already have learnt of other circumstances that might also influence our judgement. In any case, the evidence that establishes that H has occurred will itself contain, explicitly or implicitly, a wealth of further detail, which will modify our final state of information and, most likely, our probabilistic judgement.

In the Appendix (Section 16), we shall present some further critical comments relating to this topic. In any case, it should always be borne in mind when dealing with a problem of inductive reasoning (and, were it not for fear of annoying the reader, we should certainly stress this more frequently).

11.2.3. This, then, is what 'inductive reasoning' is all about. It is often said to reveal how it is that one 'learns from experience', and this is true, up to a point. It must be made clear, however, that experience can never create an opinion out of nothing. It simply provides the key to modifying an already existing opinion in the light of the new situation. The complex A (the experience) *by itself* determines nothing, nor does it provide bounds: to reach a conclusion – that is to determine a new ('posterior') opinion $\mathbf{P}$ – we require the conjunction of A with $\mathbf{P}^0$ (the initial, or 'prior', opinion). This should not be interpreted as the experience (represented by A) disproving $\mathbf{P}^0$, or forcing one to discard it in favour of $\mathbf{P}$. On the contrary, the adoption of $\mathbf{P}$ in the new state of information is the only way of remaining consistent with what was adopted as the initial opinion in the initial state of information.[8] So far as the terms 'prior' and 'posterior' are concerned, they simply signify 'before' and 'after' the acquisition of the information A. One should avoid giving too much weight to this, lest the impression is given that 'prior' refers to some mysterious circumstance of being 'prior to any experience', or to a state of 'absolute ignorance', or 'total indifference' and so on, or even that we are referring to different kinds of probability (as was the case with the old terminology relating to *a priori* and *a posteriori* probabilities).

Better still, remembering that there are two sides to every relationship, we could say that equation 11.1 merely reveals the possibility of evaluating $\mathbf{P}(E)$ in two different ways: directly, having in mind the final state of information, H_0 plus A, or by evaluating $\mathbf{P}^0(E)$ and $\mathbf{P}^0(AE)$, thinking only in terms of our previous state of information H_0. Coherence requires that the two answers be the same. If one tries it both ways and finds a difference, then the evaluations should be reconsidered, their reliability checked by each method, and adjustments made on this basis until they coincide. This is not a question of deduction (albeit within the ambit of evaluations of subjective probabilities) so much as an invitation to reflect on one's own opinions in order to make them compatible with the requirements of coherence. We point this out explicitly, largely because – for

8 Recall – or, preferably, re-read – the discussion given in Chapter 4, 4.5.3, and in Chapter 5, Section 5.9; see also Section 11.3.1 of this chapter.

obvious reasons of simplicity – our own exposition will always follow the same path; first evaluating $\mathbf{P}^0$, subsequently observing A, and, finally, arriving at the conclusion $\mathbf{P}$.[9]

11.2.4. One fact to note is that the explanation that we have given has only one form and is suitable for every application of inductive reasoning, with no exceptions. This will seem natural to those who have entered into the spirit of the subjectivistic conception of probability, and would scarcely be worth mentioning at all, were it not that certain other approaches consider *statistical induction* – usually referred to as *statistical inference* – as a case apart, and, indeed, as the only case in which probability theory finds any legitimate application.

According to these other approaches, statistical inference is the special form of reasoning to be applied when a large quantity of related data is available. For example, when the frequency of some given phenomenon in a large number of trials is known, or when we know the percentages of people in a given population who possess certain characteristics, and so on. The conclusions that are put forward on this basis derive their overall justification from the fact of there being a large quantity of data. They are valid, therefore, insofar as the quantity of data is sufficient for them to be regarded as such, and not otherwise.

To use a classical form of terminology, we would be dealing with a property connected with the existence of an 'aggregate'. So long as we are dealing with just a few objects, they do not form an aggregate and no conclusions can be reached. If, however, we have a large number of objects, then we do have an aggregate and then, and only then, does the argument go through. If we add in objects one at a time, nothing can be said until the number of objects becomes sufficient to be considered an aggregate; then the conclusion appears (just like that? in passing from 99 to 100? or from 999 to 1000? …), as that which is not yet an aggregate at last becomes one. Now it will be objected that this version is a travesty: there is no sharp break of this kind, but rather a gentle transition. The nonaggregate passes through a to-be-or-not-to-be-an-aggregate phase, inclining first one way and then the other, and only subsequently does it gradually transform itself into a real and genuine aggregate. But this does not answer the original objection raised against the distinction, here put forward as being of fundamental conceptual importance, between the 'aggregate effect' and the 'effect of individual elements'. To recognize that a clear-cut separation cannot exist, even though this admission may perhaps resolve certain of the apparent paradoxes, does not get to the real root of the problem and, indeed, serves to underline the weakness and the contradictory nature of the whole approach.

---

9 This last comment might help to reduce the impact of one rather obvious objection that springs to mind: if from $\mathbf{P}$ we require to trace back to $\mathbf{P}^0$, should we not trace back from $\mathbf{P}^0$ to some $\mathbf{P}^{00}$, and so on, *ad infinitum*? Where, then, would the very first evaluation have come from? The question is rather sophistical, since the procedure which we have given loses its force when carried out in situations that are too far removed from reality. On the other hand, it is well known – and we shall see examples of this – that even very vague prior evaluations are often sufficient to yield conclusions of more than adequate practical precision (and this holds, *a fortiori*, if we retrace from one 'beginning' to another). Eventually, perhaps, we should need to have recourse to an explanation based on 'instinct', or to experience in the form of genetic inheritance, or something of that kind (I do not want to insist too seriously on these suggestions relating to fields in which I am not expert). For a more detailed discussion, see B. de Finetti and L.J. Savage (1962).

The problem is only resolved by acknowledging that distinctions of this kind have no significance. The conclusions one arrives at on the basis of a large quantity of data are not the consequence of some aggregate effect, but simply the cumulative effects of the contributions of the individual pieces of information. The modification of a prior opinion into a posterior opinion through knowing the outcome of some given set of trials is precisely the same as that obtained by considering each item of data separately, and effecting the appropriate modifications (in general, minor) one at a time. This is so no matter whether the number of trials is large or small and is an important fact to bear in mind if serious misunderstandings are to be avoided. We are aware that we have, perhaps, given undue emphasis to this point, but the fact remains that the germs of such misunderstandings seem to permeate the very air we breathe.

11.2.5. In what follows, problems of the 'statistical type' will receive their due emphasis; they are undoubtedly interesting from a theoretical point of view, and certainly important so far as practical applications are concerned. They will, however, simply be a special case – or, more precisely, a collection of rather ill-defined special cases – having in common the following general characteristics: that past experience consists of a number of observations of 'more or less similar facts' (and often the case of interest is that of a large number of such observations). Of course, analogy per se is a rather marginal and irrelevant factor but it often leads one to considerations of some kind of symmetry in the evaluations of probability, and this is what really concerns us (even though, from a descriptive point of view, it is often useful to mention the analogy in question).

As a more expressive statement of the way in which such an analogy is translated into probabilistic terms, we could say that the analogy leads us to make the conclusions depending only, or, at least, mainly, on *how many* 'analogous' events occur out of some given number, and not on *which* of them occur. This is intended to give a broad view of what we mean, however, and should not be interpreted in any literal sense.

It is within this kind of framework that we consider the problem of evaluating probabilities on the basis of observed frequencies (and although we have touched on this topic before – see Chapter 5, Sections 5.8–5.9 and Chapter 7, 7.5.5–7.5.6 – we have not done so in a systematic fashion).

We shall soon see, however, that even in this case, the simplest, there is no unique answer. In fact, one is permitted – and, indeed, obliged – to choose any initial opinion from among those possible (and the latter will turn out to correspond to the set of functions which increase from 0 to 1 over the interval [0, 1]). Conversely, we shall also see that, starting from an examination of the same series of results, it may be natural to express opinions which involve extending the whole approach, although the qualitative features originally advanced as being characteristic of problems of the statistical type remain unchanged (or are changed in minor respects only).

11.2.6. In order to give a more concrete presentation of the various possible attitudes to the way in which a given set of results should influence us, we shall examine a particular example. Let us suppose that we have observed 50 events, E_1, E_2,..., E_{50}, and that the results are the following:

1111001111 0111000111 1100001110 0000111011 1000000010,

where 1 denotes a success, and 0 a failure.

What can we say on the basis of these results? What probability should we attribute to some other event, E_{51}, or E_{312}? Or to the proposition that there will be k successes ($k = 0, 1, 2,\ldots, 100$) out of some other collection of 100 events (either a particular, preassigned collection, or just some collection chosen, in some specified sense, 'at random')?

It does not make sense to pose these questions in this abstract fashion. We have got to know what kind of events we are dealing with, and what information we have concerning them, no matter how limited it may be. To say there is 'no available information' is too glib: were this actually the case, we would not even know what kinds of events we were dealing with (they would simply exist as E_i, $i = 1, 2,\ldots, 50$). In such a case, there could be no possibility of considering their probabilities, nor any interest in doing so. Even in this case, however, in trying to figure out why the author has presented such an example, the reader would form some opinion, albeit tentative, and it would be this opinion that was relevant, rather than the so-called state of 'no available information'.

11.2.7. In real-life examples, one will have some idea of which features or attendant circumstances might lead to different probabilities of success (in the sense that one feels inclined to treat as meaningful, and to take into future account, any significant departures from the norm in the frequencies for events possessing these features, or being dependent on these circumstances). If, for example, in considering the deaths resulting from an epidemic one finds significant differences for individuals with different blood pressures, or for those born at different times of the year, or on different days of the week, there will be a tendency to regard the differences as meaningful in the first case but not in the others.

Another circumstance, which may or may not appear as meaningful, is that of order. In our example, we assumed the events to be numbered from 1 to 50: in many cases, such a numbering is just a matter of convention and is completely irrelevant (registration numbers, passport numbers etc.), but in others it will correspond to chronological order and then may well be meaningful, in that it could reveal a tendency for the frequency of successes to increase or decrease with time, or to oscillate. Moreover, in cases where the fact of two trials being consecutive is meaningful, study of the order may reveal a difference in frequencies for trials depending on whether they follow a success or failure (and one could easily consider other variants of this idea).

In any actual example, there are innumerable different factors that could be considered in this way, the vast majority of which would certainly be meaningless. But there may be other examples in which these same factors in various combinations will appear, to some extent at least, meaningful. In any case, this rather general summary finds its genuine expression only in the evaluation of the probabilities for all the constituents constructed on the basis of the events under consideration. Alternatively, if one prefers to look at it in this way (the two approaches are equivalent), we consider the probabilities conditional on every possible combination of results out of any group of observed events, these being taken over every possible combination of other events. Within this framework, everything can be expressed in a complete form; this is true for all possible cases.

11.2.8. It is clear, however, that it would be very difficult to consider simultaneously all possible factors and, in any case, this would only cause confusion. In studying this topic from a theoretical viewpoint, therefore, one restricts oneself to considering certain relatively simple cases in which only a small number of factors (and sometimes only one) are considered. In any practical application, of course, one must not lose sight of the fact that simplified schemes of this kind are likely to be inadequate in certain respects.

We should also make it clear that the various *schemes* to which we shall make reference (those of Bernoulli, Poisson and Markov, together with the exchangeable and partially exchangeable cases, contagion models and so on) should not be interpreted as fixed slots into which real applications are to be fitted. Still less should they be viewed primarily as mathematical inventions, whose complications are merely evidence of mathematical playfulness, and which are devoid of interest so far as applications are concerned. They should be seen rather as simplified schemes serving as possible representations of the one and only realistic 'scheme' – that which includes all possible distinctions in all possible combinations. The schemes we shall deal with are useful in practice, but, again we note, only as simplified representations of more complicated situations which themselves cannot be represented straightforwardly.

11.3 The Case of Independence and the Case of Dependence

11.3.1. *Independence.* We first consider the case in which the possibility of observed results influencing subsequent evaluations is specifically excluded. This is the case of independence, with which we are already familiar, and *for which the problem of inference does not exist.* The case would not concern us, therefore, were it not for the following two considerations, the first of which is of a technical nature. It turns out that the most convenient way of attacking problems of interdependence is to reduce them, if possible, to appropriate combinations of independent schemes (as we shall see in Section 11.3.5 and subsequently). The second consideration is of a critical nature: the statement that the problem of inference does not exist in the case of independence, although obvious, often gives rise to misunderstandings (more precisely, it is misguidedly dismissed by those who have not properly understood what it actually says). What it says is that any possibility of 'learning through experience' is *excluded* – 'ruled out by the principle of contradiction' – if the original opinion is based on independence, because the latter, by definition, requires that the original opinion will not be modified on the basis of any observation of results.

Let us consider, as an example, the case of Heads and Tails, with the assumption that the two probabilities are equal (i.e. $\frac{1}{2}$) and that trials are independent. The evaluations of probabilities for successive trials remain unchanged, no matter what results are observed (like, for instance, those considered above in Section 11.2.6, supposing them to be the results of the first 50 trials).

The same could be said in the case of a die if, for example, we considered the face '1' as a success, and the others as failures, and assumed throws to be independent (we merely have $p = \frac{1}{6}$ in place of $p = \frac{1}{2}$). In the case of independence – that is if the original opinion is based on an assumption of independence – every possibility of 'learning through experience' is ruled out (it would not be consistent with the original opinion).

Someone might, perhaps, argue as follows (in the context of the example of Section 11.2.6). If the die gives me face '1' 26 times out of 50 (instead of about 8 times), I am inclined to believe that it is 'loaded' (i.e. that it favours '1', and perhaps there is a weight in the opposite face): it also happens that 18 times out of 26 '1' is followed by '1', and 16 times out of 23 '0' is followed by '0'; I suspect that the way the die is thrown favours 'repeats', and this leads me to revise my original assumption of independence

and to drop it. Moreover, noticing that the number of times '1' occurs in the five blocks of ten decreases from 8, 6, 5, 5 to 2, I am led to think that the loading which originally favoured '1' was temporary and subsequently ceased to operate, so that the die is now perfect. Perhaps, if I continue, I shall notice a number of other things!

Now it may be that arguments of this kind are acceptable in themselves (this is a matter of opinion), but it is necessary that they be formulated correctly, so as to avoid any possibility of misunderstanding. Insofar as they seem to be in conflict with the previous assertion concerning the contradiction involved in changing one's mind having assumed independence, we can deduce that either the form in which they are expressed, or the manner in which they are interpreted, is mistaken.

The mistake, in fact, is in referring to stochastic independence as if it were an 'hypothesis' which the facts can 'dispute', enabling us, and possibly obliging us, to change our minds. If we are to be able to 'change our mind', the original opinion must be expressed in a form that is compatible with such a possibility of revision. Such an opinion could, at most, be a 'first approximation' to the case of independence, in that it might, for example, consist of a mixture of evaluations, most of which correspond to the case of independence (with some preassigned p), but some of which, although having little weight, correspond to various *alternatives* (like those mentioned for the above example).

It is only by admitting such alternatives that a 'revision' can take place; and, indeed, not simply by admitting their possible existence, but rather through their actual presence in the original opinion, *which, therefore, can no longer possess the property of independence.* The so-called revision – that is the passage from the original opinion to a different subsequent opinion – takes place, in fact, as a result of outcomes that give rise to a strong likelihood for such an alternative: in other words, roughly speaking, if we suspect that the occurrence of a certain event should be attributed to an alternative explanation under which it would have a higher probability.

When discussing this topic previously (see Chapter 7, the first footnote to Section 7.5.8), we emphasized that one should speak of *suspicious* cases, rather than calling them 'strange' or 'unlikely', as is often done. The reason for this is the one we have just given, but it will be useful to provide further illustration, in order to clarify the contrast between our terminology and the terminology which we reject (not simply on the grounds that it is inappropriate but also because it leads to the construction and application of methods which have no proper foundation). We note that from a conceptual viewpoint the considerations which we have put forward hold completely generally: our detailed concentration on the Bernoulli scheme – in particular the special case of Heads and Tails – is purely for the purpose of fixing ideas.

Those who think in terms of a 'revision' – or even a 'disproof' – of the original opinion, without having in mind, or referring to, any alternatives, could not regard what has occurred as 'suspicious', since the word is meaningless unless alternative explanations are admitted. Instead, it would be described, having in mind the original opinion, as "strange', 'unlikely', 'exceptional', 'very improbable' or 'very unexpected'.

More specifically (and, for convenience, we deal with the simplest cases), the circumstances which would characterize the cases 'disproving' the original opinion would be one or other of the following:

- the *distance* from the prevision; this ties up with *hypothesis testing*, for example whether or not something belongs to a 'confidence interval' (with some given 'tail area' probability), or to the interval $m \pm 3\sigma$, or something similar (see Chapter 12, 12.6.4);

- the *small probability* of the case which has occurred;
- some observed *peculiarity*; for example, that all the 0 s come before all the 1 s, or that they alternate, or – should one happen to spot the fact! – that the binary sequence is the coding of some celebrated historical date.

These are circumstances that may turn out to be useful in practice; not in themselves, however, but rather if, and insofar as, they serve to strengthen more usual forms of 'suspicion' (like those regarding 'cheating', 'malfunctioning' etc.). With regard to 'small probabilities', one should say immediately that the whole thing is rather ambiguous. Is it to be taken as referring to the probability of the particular sequence of 50 1 s and 0 s, or to the probability of the frequency; that is of all those sequences in which a 1 occurs 26 times?[10]

To speak in terms of objective circumstances, rather than suspicions relating to other alternatives (and to make use of criteria based upon such objective circumstances), means, as usual, that one is attempting to draw conclusions on the basis of a single possibility, neglecting the necessary comparative possibilities.

Everyone is free to choose his prior opinion in whatever way he likes. The choice can only be made once, however. If I choose to base my prior opinion upon the assumption of independence, it means that I exclude, once and for all, any circumstance that might in future be pointed out to me as rather 'strange'. I refuse to consider it as a possibility; that is as something capable of modifying my opinion. If the future occurrence of this 'strange' circumstance would, in fact, lead me to suspect 'cheating' (or whatever), then I should make it clear from the very beginning that my opinion is not based on the assumption of independence, but that it accepts the dependence deriving from admitting the possible suspicion (which, although negligible at the outset, could, under certain circumstances, come to the fore). If I omit to say this, then, at best, I have expressed myself rather superficially (this might be excused, however, if I was aware what I was doing).

There would be no excuse, on the other hand, if a change of opinion was explained in a distorted fashion, by attributing it to the fact of experience having disproved the original opinion, dictating its replacement by another. Nothing can oblige one to replace one's initial opinion, nor can there be any justification for such a substitution. From a logical point of view – and, it might even be argued, from the 'moral' point of view – one would be adopting the same contradictory posture (or indulging in the same unfair subterfuge) as a person who regards himself as released from a promise to help a friend if a certain event occurs, given that the event in question has already occurred.

In order to retain the right of being influenced by experience, it will therefore be necessary to express an initial opinion differing from that of independence.

11.3.2. *Exchangeability*. Having abandoned independence, the simplest choice open to us is to continue to regard the order as irrelevant. Given n events, the probabilities $\omega_h^{(n)}$ that h of them occur ($h = 0, 1, 2, \ldots, n$) are now arbitrary[11] (and are no longer necessarily those of the binomial distribution, as in the case of independence); however, the

10 Not to mention the fact that in more complicated cases, or if one takes other circumstances into account, the 'observed result' (whatever it may be) always has an arbitrarily small probability if one describes it in a sufficiently precise way.

11 If, however, the events are at least potentially infinite in number, then there may be restrictions (see Section 11.4).

combinations of h 1 s and $n-h$ 0 s all have the same probability $\omega_h^{(n)}/\binom{n}{h}$. This is equivalent to simply saying that all products of n events have the same probability $\omega_h^{(n)}$.[12]

In this case, the events are called *exchangeable* (the reasons for the terminology being contained in what we have already said): knowledge of the n results can only have an influence through the reporting of n and the frequency (i.e. n and h). whereas any other aspect connected with order will be ignored. We shall return to this topic shortly (in Section 11.4).

It is intuitively obvious that drawings from an urn with unknown composition are exchangeable (e.g. an urn containing an unknown number of black and white balls, with the standard method of drawing with replacement). The same applies to tosses of a possibly asymmetric coin and, more generally, to all those cases that are commonly referred to as 'repeated trials with a constant but unknown probability of success.'[13] It is less obvious but nonetheless true, as we shall see, that we still have exchangeability in the case of drawings without replacement, or with double replacement (the 'contagion' model; see Chapter 10, 10.3.5).

In the example we have been considering, the only significant fact is that we had $h=26$ successes out of $n=50$ tosses. This can also be expressed by saying that the two numbers $n=50$ and $h=26$ are 'sufficient statistics' (i.e. they constitute an exhaustive summary of the data). In other words, so far as 'learning from experience' is concerned, it does not matter whether we observe the complete sequence, or whether we simply observe that $n=50$ and $h=26$; this is a consequence of the assumption of exchangeability.

11.3.3. *Partial exchangeability*. We obtain a somewhat less restrictive condition (although at the expense of some additional complication) by thinking of the events under consideration as divided into various classes (in order to fix ideas, we shall consider two classes) and of exchangeability as holding within both classes. In other words, the probability that out of $n=n'+n''$ events (n' of the first class, n'' of the second) $h=h'+h''$ occur (h' of the first class, h'' of the second) is the same, no matter how the n' and n'' events are chosen, and no matter which of them are among the h' and h'' successes. The probability of obtaining a total of h' and h'' successes is $\omega_{h',h''}^{(n',n'')}$ and the probability of them occurring for a particular, preassigned sequence of events is that just given, divided by $\binom{n'}{h'}\binom{n''}{h''}$. Obvious, trivial examples are that of exchangeability per se (for which ω depends only on $n'+n''$ and $h'+h''$) and that of independence between the classes (within each of which there is exchangeability; ω is then the product of the $\bar{\omega}_{h'}^{(n')}$ and $\bar{\bar{\omega}}_{h''}^{(n'')}$ for the two cases). Actual cases of partial exchangeability fall into an intermediate category: put in a rather imprecise form, but one which conveys the general idea, we have interdependence between all the events but a rather stricter one among those in the same class. This would be true, for example, of drug trials carried out on patients of both sexes.

---

12 In fact, it suffices to observe from equation 3.11 of Chapter 3, 3.8.4 that we have

$$\omega_h^{(n)} = \binom{n}{h}\sum_{r=0}^{n-h}(-1)^r\binom{n-h}{r}\omega_{h+r}^{(h+r)} = \binom{n}{h}\Delta^{n-h}\omega_h^{(h)}. \tag{11.2}$$

13 The terminology is incorrect (see Chapter 4, 4.8.3–4.8.4), but is expressive (and the meaning it suggests is essentially correct).

In the particular case of events occurring in chronological order, the division into classes may depend on the result of the previous trial; in such a case we have the Markov form of partial exchangeability.

If one suspects that an outcome is influenced by the preceding result, then one would not initially regard all sequences having the same numbers of successes and failures as equally likely (in the example, this would be those with 26 1 s, and 24 0 s). Instead, the judgement of equal probability would apply to all those sequences having the same number of successes and failures following the occurrence of a 1 (18 and 8, respectively) and the same number (7 and 16, respectively) following the occurrence of a 0. We might expect some similarity with the case of a Markov chain with probability 18/(18 + 8) (about 70%) of a success following a success, and 7/(7 + 16) (about 30%) of a success following a failure … (however, there are various reservations, as in the previous case, and these become more serious the more complicated the situation becomes). In any case, with the above assumption one requires n', n'', h' and h'' for an exhaustive summary (n and h alone no longer suffice).

11.3.4. *Other cases.* Similar conclusions hold in the other case we mentioned in Section 11.2.7; that in which one suspects a progressive increase in one of the two probabilities at the expense of the other (right from the beginning; recall that no suspicion can arise if it is not present initially). For example, we might suspect that under certain circumstances (for instance, a black ball being drawn) white balls may turn into black balls during a series of drawings with replacement from an urn of unknown composition. A study of the outcomes provides information concerning the composition of the urn if we consider the tendency for the frequency of white balls to decrease with time (this frequency being the only thing we can get hold of). It also provides a basis for making conjectures about the past history – and hence about the future – of the unobservable process by which white balls are gradually turned into black balls.

The general case follows along the same lines as all the examples which we have considered. Given an arbitrary prior probability distribution $\mathbf{P}^0$, which attributes probability to each $A = E_1'E_2' \ldots E_n'$ (where E_i' stands for either E_i or $\tilde{E}_i$, and n is arbitrary), the problem is solved by simply stating that, knowing A, the (posterior) probabilities are given by $\mathbf{P}$, where

$$\mathbf{P}(E) = \mathbf{P}^0(EA) / \mathbf{P}^0(A).$$

The extreme simplicity of this mathematical statement is, however, misleading. In general, straightforward application of the method is precluded by the requirement that one provide the $\mathbf{P}^0(A)$ directly for all A. This is only really feasible if the situation can be represented in terms of simple formulae.

11.3.5. *Mixtures of distributions which assume independence.* The straightforward case of independence is itself uninteresting; we have, simply, $\mathbf{P}^0(EA) = \mathbf{P}^0(E)\mathbf{P}^0(A)$, and hence $\mathbf{P}(E) = \mathbf{P}^0(E)$ for all E which are defined in terms of 'future' trials (or, at least, do not depend on the observations A). As we mentioned in Section 11.3.1, however, it turns out that in a number of cases it is extremely useful to consider the possibility of expressing $\mathbf{P}^0$ as a *mixture* of such distributions $\mathbf{P}_i$: in other words, we take linear combinations

$$\mathbf{P}^0 = c_1^0\mathbf{P}_1 + c_2^0\mathbf{P}_2 + \ldots + c_m^0\mathbf{P}_m = \sum_{i=1}^{m} c_i^0\mathbf{P}_i,$$

with non-negative coefficients c_i^0 having sum equal to 1 (or limit cases thereof). The latter may take the form of infinite series, or of integrals; in either case, it is sufficient to describe it as a $\mathbf{P}^*$ for which, given any arbitrary $\varepsilon > 0$, there exist $\mathbf{P}$s having the form of finite linear combinations such that $\sup_E |\mathbf{P}^*(E) - \mathbf{P}(E)| < \varepsilon$.

It is easily seen that if $\mathbf{P}^0$ is a mixture, then so is any $\mathbf{P}$ to which it leads as a result of (arbitrary) observations A (the coefficients varying for different experiences, A). In fact, we have

$$\begin{aligned}\mathbf{P}(E) &= \frac{\mathbf{P}^0(EA)}{\mathbf{P}^0(A)} = \frac{c_1^0\mathbf{P}_1(A)\mathbf{P}_1(E) + \ldots + c_m^0\mathbf{P}_m(A)\mathbf{P}_m(E)}{c_1^0\mathbf{P}_1(A) + \ldots + c_{\boldsymbol{m}}^0\mathbf{P}_m(A)} \\ &= c_1\mathbf{P}_1(E) + \ldots + c_m\mathbf{P}_m(E),\end{aligned} \tag{11.3}$$

where $c_i = Kc_i^0\mathbf{P}_i(A)$ (K = the normalization factor = $1 / \sum c_i^0\mathbf{P}_i(A)$).

The expression in mixture form may correspond to an actual mixture, in which case there exist events, $H_1, H_2, \ldots, H_m$ (exclusive and exhaustive) such that the $\mathbf{P}_i$ represent probability distributions conditional on the H_i: $\mathbf{P}_i(E) = \mathbf{P}(E|H_i)$. In other cases, where this does not apply, it may, nevertheless, turn out to be useful to proceed, formally, *as if* such events existed.

11.3.6. In the *exchangeable* case, if we think in terms of an urn of unknown composition, the H_i represent the events (or 'hypotheses') that the proportion of white balls is θ_i (and under such an hypothesis we consider the drawings to be independent and of constant probability, $p_i = \theta_i$). If we think in terms of a biased coin, or of the Pólya urn scheme, objective circumstances of this kind (i.e. observable in principle, even though we cannot actually observe them) do not exist. However, we shall soon see (in Section 11.4.2) that, in the exchangeable case, $\mathbf{P}^0$ always has the form of a mixture. This will then permit us to argue as if the coefficients c_i^0 and c_i were the probabilities of events H_i, conditional on which we have independence and probability of success equal to p_i.

11.3.7. In the *Markov* case (dependence on the preceding result), we can still reduce to mixtures by considering distributions $\mathbf{P}_i$ under which the trials are independent with probability p_i' or p_i'', depending on the outcome of the previous trial.

In the third example (that of decreasing probabilities), it may or may not be possible to reduce to a mixture, depending on the way the initial opinion is stated. The statement considered previously does not permit us to do this. On the other hand, we do have a mixture of distributions if the latter are taken to be of the form

$$\mathbf{P}_i(E_h) = f_i(h),$$

where the E_h are independent according to the $\mathbf{P}_i$, and the $f_i(h)$ are arbitrary (e.g. $e^{-\lambda_i h}$, if one requires them to be decreasing).

11.4 Exchangeability

11.4.1. We shall now consider the general notion of exchangeability and, in particular, exchangeable events and exchangeable random quantities. What we are considering, in fact, is the most fundamental and widely used form of statistical inference.

The definition of exchangeability in the case of events has already been given, but we shall re-express it in such a way as to include, also, the case of exchangeable random quantities. The definition is the following: for arbitrary n, the distribution function, $F(.,.,\ldots,.)$, of $X_{h_1}, X_{h_2}, \ldots, X_{h_n}$ is the same, no matter how the X_{h_i} are chosen (in particular, F must be symmetric, because the X_{h_i} could simply be permuted). More generally, every condition concerning n of the X_h has the same probability, no matter how the X_h are chosen or labelled.

We shall come across applications of exchangeable (and partially exchangeable) random quantities in Chapter 12. For the time being, we shall restrict ourselves to establishing a particular property that we shall make use of in the special case of exchangeable events (a topic to which we shall return shortly).

Let us consider exchangeable X_h having *finite variances* and, in particular, we shall look at two large groups of such quantities. What we shall prove, roughly speaking, is that their arithmetic means, Y'/n' and Y''/n'', are almost certainly equal (where Y' and Y'' are the sums of the n' quantities in the first group and the n'' in the second, respectively). More precisely, we shall show that the square of their difference tends to zero in prevision as n' and n'' increase. This gives us Cauchy convergence in mean-square, and hence weak convergence, and a limit distribution F for the mean Y_n/n of a large number of terms: $F_n \to F$, where $F_n(x) = \mathbf{P}(Y_n/n \leqslant x)$.

The proof is as follows (and is given under conditions that are less restrictive than those of exchangeability). We assume that the X_h have the same, finite, previsions and variances, m and σ^2, and the same pairwise correlation coefficient, r.[14]

Expanding the square of $n''Y' - n'Y''$, we obtain $n'n''(n' + n'')$ terms of the form X_hX_k with $h = k$, and the same number[15] (but with the opposite sign) having $h \neq k$. The previsions are $m^2 + \sigma^2$ and $-(m^2 + r\sigma^2)$, respectively, so that

$$\mathbf{P}\left(\frac{Y'}{n'} - \frac{Y''}{n''}\right)^2 = \frac{n' + n''}{n'n''}\left[\left(m^2 + \sigma^2\right) - \left(m^2 + r\sigma^2\right)\right] = \left(\frac{1}{n'} + \frac{1}{n''}\right)\sigma^2\left(1 - r\right). \tag{11.4}$$

We could have set $m = 0$ at the very outset (it disappears in the formulation of the problem), but it is sometimes useful to have the formula available for $m \neq 0$. This is particularly so in the case of exchangeable events, because $\mathbf{P}\left(E_h^2\right) = \mathbf{P}\left(E_h\right) = \omega_1$ and $\mathbf{P}(E_hE_k) = \omega_2 (h \neq k)$, and hence

$$\mathbf{P}\left(\frac{Y'}{n'} - \frac{Y''}{n''}\right) = \left(\frac{1}{n'} + \frac{1}{n''}\right)\left(\omega_1 - \omega_2\right). \tag{11.4'}$$

---

14 We remind the reader that if (at least in principle) the X_h are infinite in number then $r \geqslant 0$ (see Chapter 4, 4.17.5). So far as we are concerned, the X_h must be infinite in number – or, at least, very numerous – and so r will always be positive – or, at worst, negative but very small (and $1 - r$ will be $\leqslant 1$, or just greater than 1).

15 The 'number of terms' is to be understood in an algebraic sense (i.e. counted as -1 if it has a sign opposite to that being understood).

Note that the two groups are assumed to be disjoint: if they had c terms in common, we have a tighter bound (as one might have expected). The factor $n' + n''$ becomes $n' + n'' - 2c$ (i.e. $2c/n'n''$ is subtracted from $(1/n') + (1/n'')$).

Returning now to the case of exchangeable events and thinking of them as a sequence (but one whose ordering is arbitrary and irrelevant), we can characterize them as a stochastic process with the same representation we used for Heads and Tails: all paths leading from the origin to some given point are equally probable. In this case, we shall speak of an *exchangeable process.*

For such a property to hold, it is sufficient that the probabilities $p_h^{(n)}$ and $\tilde{p}_h^{(n)}$ of steps of +1 or −1, respectively, when leaving a given point $[n, h]$,[16] depend on the vertex in question, but not on the path travelled in order to reach it, and that the probability of successive steps of +1 and −1 remains unchanged if they are reversed: that is $p_h^{(n)}\tilde{p}_{h+1}^{(n+1)} = \tilde{p}_h^{(n)}p_{h+1}^{(n+1)}$. This condition lends itself to an elegant geometrical interpretation. If, at each vertex, the probabilities of the next step are expressed as a vector $(1, p - \tilde{p})$ (where $p = p_h^{(n)}$) pointing at the barycentre (or prevision) of the possible points of arrival, then the condition may be stated as follows: for any vertex $[n, h]$ and the two following, $[n+1, h]$ and $[n+1, h+1]$, the three corresponding vectors *meet at a point* (see Figure 7.2 of Chapter 7, 7.3.3). By induction, this condition is itself sufficient to ensure exchangeability (provided it holds at all vertices of the lattice).

As a function of the ω, we have

$$p_h^{(n)} = \frac{h+1}{n+1}\left(\omega_{h+1}^{(n+1)}/\omega_h^{(n)}\right). \tag{11.5}$$

Note that each path from the origin to $[n+1, h+1]$ has probability $\omega_{h+1}^{(n+1)}/\binom{n+1}{h+1}$, whereas those paths coming from $[n, h]$ have probability $\left[\omega_h^{(n)}/\binom{n}{h}\right].p_h^{(n)}$. Equation 11.5 follows on comparing the two probabilities.

11.4.2. Some processes necessarily come to an end in a finite number of steps (for example, drawings without replacement from an urn containing N balls; if H are white, $0 < H < N$, we have $\omega_H^{(N)} = 1$, and so we cannot continue with the $\omega_h^{(n)}$ for $n > N$): others can be considered as if they could be continued indefinitely.

All exchangeable processes that end after N steps are mixtures of the *hypergeometric* process. The mixtures over the possible cases $H = 0, 1,\ldots, N$, with probabilities c_0, $c_1,\ldots, c_N$ (H unknown, and perhaps chosen at random in some way or other), coincide, within the N steps, with every process that has, for $t = N$, the given distribution for Y_N: that is

$$\mathbf{P}\left(Y_N = 2h - N\right) = \mathbf{P}\left(S_N = h\right) = \omega_h^{(N)} = c_h \quad \left(h = 0,1,\ldots,N\right).$$

The idea is obvious: if the probabilities of passing through the various vertices of the vertical line $t = N$ are made to coincide (by balancing the drawing), then, for any two process, all the probabilities relating to occurrences before time t (displayed on the left in the figure) also coincide. These latter probabilities are all well determined, since all the paths ending at a given point have equal probabilities.

16 See Figure 7.1 in Chapter 7, 7.3.2.

More important, however, is the case of exchangeable processes, which can be continued indefinitely. Clearly, those obtained by mixtures of Bernoulli processes, that is such that

$$\omega_h^{(n)} = \int_0^1 \binom{n}{h} \theta^h (1-\theta)^{n-h} \, \mathrm{d}F(\theta), \tag{11.6}$$

are of this type. In the discrete case, or if a density exists, we have

$$\omega_h^{(n)} = \sum_{i=1}^{m} c_i \binom{n}{h} \theta_i^h (1-\theta_i)^{n-h} \tag{11.6′}$$

and

$$\omega_h^{(n)} = \int_0^1 \binom{n}{h} \theta^h (1-\theta)^{n-h} f(\theta) \mathrm{d}\theta. \tag{11.6″}$$

Conversely, it can be shown that every exchangeable process which can be continued indefinitely is a mixture of this form. In order to prove this, it is sufficient to refer to the previous case. For any given N, we know how to construct an exchangeable process coinciding (in $0 \leqslant t \leqslant N$) with our given process. As we have seen, this can be achieved as a mixture of hypergeometric processes with N steps: it suffices that the urn having composition H/N (H white balls out of N) be chosen with the same probability as is attributed to the frequency H/N ($= S_N/N$) in the given process. If

$$F_N(\theta) = \mathbf{P}(S_N/N \leqslant \theta)$$

denotes the distribution function of the frequency in N trials, and Ber(N, θ) and hyp(N, θ) are used to denote, symbolically,[17] the Bernoulli and hypergeometric processes that result from drawing *with* and *without* replacement, respectively, from an urn containing N balls, $H = N\theta$ of which are white, then our process is given by the mixture

$$\int_0^1 \mathrm{hyp}(N, \theta) \, \mathrm{d}F_N(\theta). \tag{11.7}$$

But we know that, as $N \to \infty$, hyp(N, θ) $\to$ Ber(N, θ) and $F_N \to F$ (see Section 11.4.1), so that, in the limit, the form given in equation 11.7 tends to that of equation 11.6. There would be no difficulty in providing a rigorous treatment but it seems more instructive to emphasize the basic idea and to give an intuitive understanding of the general validity of the mixture form (and in doing so, we have opened up the way for a rigorous proof).[18]

---

17 We are not dealing with an abstraction, but rather with a convention of notation for indicating that in place of hyp(N, θ) we could put $\omega_h^{(n)}$, or any other probability (or prevision), whose value in our process will be given as a mixture by equation 11.7.

18 It can happen that by following through a sequence of logical steps one is forced willy-nilly to concede the truth of something without ever seeing what Federigo Enriques used to call the *wherefore*. I happen to believe that the wherefore is all important (a point I have repeatedly emphasized, and do not wish to dwell upon here). See, e.g., B. de Finetti, 'Sulla suddivisione casuale di un intervallo: spunti per riflessioni', in '*Rend. Sem. Mat. e Fis.*', **XXXVII**, Milan (1967) (especially numbers 1, 2, 5 and 6).

This representation in mixture form enables us to obtain, in the way we have indicated, the modified distribution resulting from the knowledge of some given number of trials, yielding r successes and s failures, say. We find that $F(\xi)$ must be replaced by $\bar{F}(\xi)$, where

$$\mathrm{d}\bar{F}(\xi) = K\xi^r(1-\xi)^s\,\mathrm{d}F(\xi). \tag{11.8}$$

We still have a process consisting of exchangeable events, but now with a probability distribution modified in proportion to the likelihood, $\xi^r(1-\xi)^s$. In other words, proportional to the ξ and $(1-\xi)$ deriving from the effect of each success and failure, respectively.

In particular, the probability for each individual trial, which is given initially by $\omega_1^{(1)} = \int \xi\,\mathrm{d}F(\xi)$ (i.e. by the abscissa of the barycentre of the distribution F), becomes, similarly, after r successes and s failures, the barycentre of $\bar{F}$: that is to say,

$$p_r^{(r+s)} = \int \xi.\xi^r(1-\xi)^s\,\mathrm{d}F(\xi). \tag{11.9}$$

11.4.3. We shall give the details for a very simple special case: that for which the initial distribution is uniform ($f(\xi) = 1$ $(0 \leqslant \xi \leqslant 1)$, $F(\xi) = \xi$). This is the classical Bayes–Laplace version, which corresponds to the idea that 'knowing nothing about the probability' obliges one to assume the uniform distribution as the 'probability of the unknown probability'. We do not regard the uniform distribution as having any special status, and still less do we subscribe to these kinds of underlying assumptions; indeed, we regard them as meaningless and metaphysical in character. On the other hand, there is some value in considering a simple, clear example; especially one which provides us with an opportunity to make some useful points. We have, in fact, already mentioned this case (in Chapter 10, 10.3.5 and 10.4.1) in relation to the problem of subdividing an interval, and in connection with Pólya's urn scheme for contagion models.

In the subdivision of the interval [0, 1], the division point chosen first, P_0, has, like any other division point, a uniform distribution. Knowing its position ξ, the event that any particular one of the other division points $P_1, P_2, \ldots, P_n, \ldots$ falls to the left of P_0 will have probability ξ, independently of the others. If ξ is not known, the probability that h out of some other n division points fall to the left of P_0 is $1/(n+1)$ for every h (i.e. all the frequencies are equally probable), because P_0 is equally likely to be any one of the $n+1$ ordered division points 'chosen at random'. If we assume that we know there to be r out of n points to the left of P_0, then the probability of a success with the next division point (i.e. that P_{n+1} falls to the left of P_0) is given by $(r+1)/(n+2)$, because the $n+1$ points divide the interval into $n+2$ pieces, $r+1$ of which are to the left of P_0 (and all have exactly the same probability of containing the new division point – assuming that nothing is known about their lengths, etc). The probability distribution of P_0, which is initially uniform, is no longer such if we know that out of a further n 'random' subdivision points r have fallen to the left of P_0. It is, instead, the beta distribution $f(\xi) = K\xi^r(1-\xi)^{n-r}$, because P_0 is then the $(r+1)$st point from the left out of $n+1$ 'random' points.

In this way, we have again displayed the likelihood factors, the equal probabilities of the frequencies, and also the value of the probability after observing r successes out

of n trials. In other words, we have found the barycentre of the beta distribution without evaluating the integral (which, in any case, would give the same result):

$$K\int_0^1 \xi.\xi^r\left(1-\xi\right)^{n-r}\,\mathrm{d}\xi = \left(r+1\right)/\left(n+2\right) \qquad \left(K=\left[\int_0^1 \xi^r\left(1-\xi\right)^{n-r}\,\mathrm{d}\xi\right]^{-1}\right). \tag{11.10}$$

This result can be expressed more appealingly by saying that, in the Bayes–Laplace case, the probability for any future trial is given by the observed frequency, modified by adding in two fictitious observations, one a success, the other a failure. This is Laplace's celebrated 'rule of succession'.

11.4.4. The same rule reveals, on the other hand, the identity of the Bayes–Laplace scheme and that of Pólya's contagion model. In the latter, in fact, one adds to two initial balls, one white and the other black, as many white and black balls as there have been draw from the urn (the result of double replacement). After n drawings, r of which resulted in the drawing of a white ball, we shall have $n+2$ balls in the urn, $r+1$ of which are white. The probability of drawing a white ball is then $(r+1)/(n+2)$.

This establishes that the probabilities are all identical to those of the previous case: in particular, the drawings are exchangeable events, the frequencies (out of a given number of drawings) are equally probable and so on. It follows that not only is 7/10(= (6+1)/(8+2)) the probability of drawing a white ball at the 9th drawing after six of the previous eight have resulted in white (in which case, we know that at this point the urn contains 10 balls, seven of which are white), but it is also the probability of drawing a white ball on any occasion for which we do not know the outcome, provided that, out of eight observed drawings, six (for instance, the 3rd, 8th, 19th, 52nd, 53rd, 100th) resulted in white balls and two in black (the 1st and 92nd, say). And this will hold for the 2nd drawing (even though it is certain that at that moment there were three balls in the urn, one of which was white), the 4th (even though there were then five balls in the urn, either two or three of which were white), the 20th, 50th, 200th or 1000th, or any other (although in determining the proportion of white and black balls the need for information becomes less and less). At the second drawing, if I only knew the outcome of the first (black), I would attribute a probability of $\frac{1}{3}$ to white, there being certainly one white and two black balls in the urn. Although this is clear-cut, the knowledge of the subsequent outcomes leads me to attribute a probability of $\frac{7}{10}$ to obtaining a white ball in that same drawing. This is because the subsequent prominence of white balls leads me to assume that their percentage increased due to a number of drawings of white balls – including, perhaps, on the second drawing.

The resolution of what appeared at first sight to be a paradox is instructive, because it makes one aware of the traps that one can so easily fall into. In this way, one's attention is drawn to the kinds of misunderstanding that may persist (due, in part, to an inability to rid oneself of past habits), even without one noticing, and even if one has thought carefully about what we have said so far, and has made an effort to adjust to our perspective and terminology. We have stated repeatedly that probability can only mean probability as evaluated by someone on the basis of available information. In this sense, the Bayes–Laplace and Pólya schemes are identical, because anyone who adopts a given prior probability distribution and has the same information (concerning the outcomes of certain events in the scheme) must evaluate the probabilities in the same way.

There may, however, be a temptation to regard these probabilities of ours as less concrete or less valid than other things that might more justifiably be called true probabilities: for example, the actual and unchanging composition of the urn in the first case, or any of the momentary compositions in the ever-changing Pólya scheme. On the contrary, these other things are either irrelevant, or even illusory. The composition of the urn (in the Bayes–Laplace sense) does make sense if we are actually dealing with drawings from an urn and is connected with the idea of probability conditional on the knowledge of such a composition. But this is irrelevant, because it is assumed that we do not have knowledge of the composition of the urn. Nevertheless, it may serve to highlight the interpretation of the distribution as a mixture.

On the other hand, *to posit an imaginary urn for the purpose of giving a more concrete interpretation to the expression in mixture form, and to the symbols in that expression in mixture form, and to the symbols in that expression which replace the probability*, and then, in this context, to refer to the 'true, unknown probabilities', is a distortion that leads to an immediate confusion of the issues. It would be equally illusory, and just as much a distortion, to imagine that behind every set of exchangeable events with an initial distribution judged to be uniform, there exists, or can be assumed to exist, a Pólya scheme whose probabilities, from drawing to drawing, are to be interpreted as a composition obtained as a result of drawing with double replacement. We have seen that changing the order changes everything, even when the above scheme actually exists.

The *one genuine and real factor* is the *probability* (albeit subjective and relative to the person making the evaluation – and, indeed, precisely because of this) that one evaluates in the actual situation pertaining (and in future situations, with respect to certain hypothetical and as yet unavailable information, which will subsequently be obtained). If we step out of this ambit, we not only find ourselves unable to reach out to something more concrete, but we tumble into an abyss, an illusory and metaphysical kingdom, peopled by Platonic shadows.

11.4.5. The considerations we have put forward in the preceding sections should be carefully studied. Not only do they provide the necessary basis for a valid conceptual approach, but they also serve to give one a clear practical awareness of how, under conditions like those which characterize the case of exchangeable events, one can justify evaluating probabilities on the basis of observed frequencies for events that are, in some sense, 'similar'. The safest and most down-to-earth approach consists, as always, in confining attention to just those particular events which are of interest to us, and, within this framework, considering the smallest number possible (without positing any infinite sequences, or any imaginary, fictitious underlying schemes). For example, if we have observed r successes and $n-r$ failures, then, in the exchangeable case, the probability which we attribute to a success on any other trial is given by

$$p_r^{(n)} = \left[\frac{\omega_{r+1}^{(n+1)}}{\binom{n+1}{r+1}}\right] \Big/ \left[\frac{\omega_r^{(n)}}{\binom{n}{r}}\right] = \frac{r+1}{n+2} \Big/ \left[1+\left(1-\frac{r+1}{n+2}\right)\left(\frac{\omega_r^{(n+1)}}{\omega_{r+1}^{(n+1)}}-1\right)\right] \tag{11.11}$$

(as is easily verified). This shows that, provided the probabilities attributed initially to the two frequencies $r/(n+1)$ and $(r+1)/(n+1)$ out of $n+1$ trials do not differ greatly, this probability itself differs little from the frequency (or the modified frequency, as in the Bayes–Laplace scheme).

11.4.6. If one uses the properties of the likelihood and the mixture form, a stronger conclusion can be obtained, although somewhat indirectly. After n trials, r of which are successes, the function $\xi^r(1-\xi)^{n-r}$, which represents the likelihood (and, in the Bayes–Laplace case, the density), increases in the range 0 to $\xi = r/n$, where it attains its maximum, and then decreases as we move from r/n to 1. It vanishes at the end-points 0 and 1 (provided, of course, that $0 < r < n$) and, if r and $n-r$ are large, it is practically 0 everywhere except in the immediate neighbourhood of the maximum. We can see this clearly by observing that, as n increases with $r/n = \bar{\xi}$ held fixed, one obtains, in the limit, the density function of the *normal* distribution, centred at the frequency, $\bar{\xi} = r/n$, and having standard deviation

$$\sqrt{\left[\bar{\xi}\left(1-\bar{\xi}\right)/n\right]}$$

(i.e. the same standard deviation that we have for the difference between the frequency and the probability for n events having constant probability ξ).

In fact, setting $x = \left(\xi - \bar{\xi}\right)/\sqrt{\left[\bar{\xi}\left(1-\bar{\xi}\right)/n\right]}$, we have, asymptotically,

$$\begin{aligned} K\xi^h\left(1-\xi\right)^{n-h} &= K\left[\xi^{\bar{\xi}}\left(1-\xi\right)^{1-\bar{\xi}}\right]^n \\ &= K\left(1+\frac{x}{\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}\right)^{n\bar{\xi}} \\ &\quad\times\left(1-\frac{x}{1-\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}\right)^{n\left(1-\bar{\xi}\right)} \to K\,\mathrm{e}^{-x^2/2}. \end{aligned}$$

To prove this, all we need to do is to take the logarithm of the penultimate expression.[19] Omitting the constant K, we obtain

$$\begin{aligned} &n\bar{\xi}\log\left(1+\frac{x}{\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}\right)+n\left(1-\bar{\xi}\right)\log\left(1-\frac{x}{1-\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}\right) \\ &= n\bar{\xi}\left[\frac{x}{\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}-\frac{1}{2}\frac{x^2}{\bar{\xi}^2}\bar{\xi}\left(1-\bar{\xi}\right)/n+O\left(n^{-\frac{3}{2}}\right)\right] \\ &\quad+n\left(1-\bar{\xi}\right)\left[-\frac{x}{1-\bar{\xi}}\sqrt{\left\{\bar{\xi}\left(1-\bar{\xi}\right)/n\right\}}-\frac{1}{2}\frac{x^2}{\left(1-\bar{\xi}\right)^2}\bar{\xi}\left(1-\bar{\xi}\right)/n+O\left(n^{-\frac{3}{2}}\right)\right] \\ &= -\frac{1}{2}x^2\left[\left(1-\bar{\xi}\right)+\bar{\xi}\right]+O\left(n^{-\frac{1}{2}}\right)\to-\frac{1}{2}x^2. \end{aligned}$$

This establishes directly that, in the Bayes–Laplace case, the posterior distribution (which is a beta distribution, with observed frequency $\bar{\xi}$ and very large n) is asymptotically normal. This conclusion holds more generally, provided that the limit distribution $F(x)$ obeys certain qualitative conditions. More precisely, it is sufficient that a density exists and is 'practically constant' in the neighbourhood of $\xi = \bar{\xi}$, and that it is not too

19 There is no mystery in the disappearance of the factor in square brackets: it does not involve x, which is the only thing we are interested in, and we have subsumed it in K.

small in comparison with distant masses, which, if they were very large, would otherwise give an appreciable contribution to the product, even though multiplied by the likelihood factor, which itself would be quite small. Such a condition – which we prefer to express in this rather vague form, because we are interested in ensuring a good approximation for large n, rather than for the asymptotic case of $n \to \infty$ – can be summarized (following L.J. Savage) by saying that the distribution F must be *diffuse* (in the neighbourhood of the point of interest).

11.4.7. The same argument also applies in cases where the scope is much wider (like those involving exchangeable random quantities rather than events), and it can be proved that the normal distribution arises quite naturally, and under relatively weak conditions even in these cases. The cases we are discussing are, of course, very different from those involving the limiting normal distribution that we discussed previously. There we were dealing with the distribution of a quantity defined as a function of a large number of other independent quantities (in particular, as the sum, but also in other ways); here, we are dealing with the form of the posterior distribution after a large number of items of information have been acquired.

From a conceptual point of view, the reason for the appearance of the normal distribution is clear (albeit in outline form) if one thinks of the genesis of the beta function in the example we have considered. It arose as the product of a number of terms ξ and $1-\xi$, each of which was the likelihood factor corresponding to an observation (the outcome of an event). In the case of observations of random quantities, also (e.g. performing a measurement, which is affected by error, of the quantity in which we are interested, or of others of which it is a function etc.), under similar conditions the likelihood factor for the totality of such observations will be the product of the factors corresponding to individual observations:

$$\nu(\xi) = \nu_1(\xi)\nu_2(\xi)\ldots\nu_n(\xi).$$

Let $\bar{\xi}$ denote the 'maximum likelihood' point: that is the point at which $\nu(\xi)$ has an absolute maximum (which we shall assume to be unique; and we further assume that $\nu(\xi)$ is much less than $\nu(\bar{\xi})$, except in a neighbourhood of $\bar{\xi}$ small enough for whatever purposes we have in mind). If, in the neighbourhood of $\bar{\xi}$, we replace $\nu_i(\xi)$ by the linear approximation

$$\nu_i(\bar{\xi}) + \nu_i'(\bar{\xi})(\xi - \bar{\xi}) = K\left[1 + a_i(\xi - \bar{\xi})\right]$$

(with $\sum a_i = 0$, in order that $\nu'(\bar{\xi}) = 0$), the product can be replaced by a polynomial for which we can repeat essentially the same form of argument as we used for the case of the beta.[20] By including with the $\nu_i(\xi)$ the factor $f(\xi)$ = prior density, the product becomes

---

20 In order to obtain the exact value of $\nu''(\bar{\xi})$, one must take into account $\nu''(\xi)$, by writing

$$\nu_i(\xi) \simeq K\left[1 + a_i(\xi - \bar{\xi}) + \frac{1}{2}b_i(\xi - \bar{\xi})^2\right].$$

This then gives the approximation

$$\nu(\xi) = K\left[1 - \frac{1}{2}(\xi - \bar{\xi})^2 \sum_i (a_i^2 - b_i)\right].$$

(Note that $\sum_{i \neq j} a_i a_j = \sum_i a_i \sum_j a_j - \sum a_i^2$).

the posterior density and the conclusions can be applied to it. If the contribution of this factor is irrelevant when compared with the others, then $\nu(\xi)$ alone already provides an approximation to the posterior density. Both the latter and the likelihood have, asymptotically, the form of the normal density.

11.4.8. Similar considerations can be made in cases where something less restrictive than exchangeability is assumed (as in those cases which we pointed out for the sake of giving examples in Section 11.3.3). In some cases the conclusions are rather similar; in others they are markedly different. The starting point and the basic ideas remain the same, however, and are always clear and straightforward. We have merely to apply to these various cases the theorem of compound probabilities, or, more directly, Bayes's theorem, which can be expressed simply in the form:

$$\text{posterior probability} = \text{constant} \times \text{prior probability} \times \text{likelihood}.$$

We should point out that many of the methods used in statistics for purposes similar to those we have been considering do not follow the lines we have indicated. They are based upon a very different set of underlying concepts and we shall not make use of them, nor shall we advocate their use. They will, however, be mentioned in Chapter 12, which is devoted specifically to statistical applications, and where it will obviously be necessary to examine and compare a number of different viewpoints and the methods they give rise to, especially those which are widely used in practice. Above all, it will be important to discuss the question of whether, and to what extent, those methods which have been introduced and justified on the basis of approaches which we consider invalid (i.e. non-Bayesian) can, in fact, be seen as legitimate (i.e. Bayesian) by suitably reinterpreting their underlying assumptions.

12

Mathematical Statistics

12.1 The Scope and Limits of the Treatment

12.1.1. A brief account of mathematical statistics within the confines of the final chapter of this book must necessarily offer a limited perspective. Nevertheless, its inclusion serves a very definite purpose.

In Chapter 11, we have already encountered certain of the problems that fall within the purview of the subject matter of mathematical statistics. Specifically, we examined applications of inductive reasoning based on statistical data; that is, data involving a number of observations (possibly a large number) that are, in a certain sense, similar to one another. We have also explained the Bayesian approach to such problems (an approach which constitutes an integral part of the subjectivistic conception), noting that a unified coherent structure cannot be maintained if this is abandoned in favour of other approaches involving a variety of more or less empirical 'ad hoc' methods.

In this chapter, we shall attempt to give a more explicit account of the problems with which mathematical statistics is concerned, and of the implications, both for these problems and more generally, which flow from the adoption of one or other of the competing points of view.

12.1.2. In addition to the strictly probabilistic aspects, with which the previous considerations are concerned, we shall have occasion to examine other topics relating to decision theory. There are two basic reasons for this and they correspond to two different questions that can be posed within the same framework.

The first of these concerns applications where the entire enterprise is more or less explicitly geared to arriving at a decision. Obvious examples of this are batch testing, or quality control, performed on a sample of a certain product in order to decide what to do with the rest of the stock (whether or not to reject it), or how to produce it in an optimal fashion (whether or not process parameters need adjusting) and so on. Although one does not always have such immediate actions in mind, it could be said that there is always some practical purpose for which one is somehow seeking guidance.

The second reason may or may not be relevant, depending on the particular application: more precisely, it depends on whether or not one is able to experiment. If this is possible – that is if one has certain choices regarding the way in which observations are to be obtained – then there is an additional dimension to the problem, because this choice is itself a decision and must be made in the most appropriate way. What is

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

appropriate will in this case, of course, depend on the final decision, itself dependent on the outcome of the experiment. More precisely, the whole question of what is appropriate needs to be set in the context of the theory of compound decisions; a framework including both decisions concerning the experiment to be performed and the final decision to be taken after the results of the experiment are known.

12.1.3. General issues will be dealt with in a somewhat summary fashion in what follows, and most of the discussion will centre around specific examples. In this way, we hope to provide a straightforward account that will make best use of the limited space available to use. The examples will, in fact, involve some of the most important and commonly occurring cases, and so, in a sense, they do provide a general perspective.

12.2 Some Preliminary Remarks

12.2.1. The cases that are usually considered in mathematical statistics are those which can, in various ways, and in a somewhat loose sense, be regarded as generalizations of the case of exchangeable events. This is in the nature of an aside, however, and simply provides a convenient reference to, and reminder of, the contents of the previous chapter: in fact, in the standard terminology of mathematical statistics one never comes across any reference to exchangeability. In order not to introduce a further difficulty into the task of comparing the various viewpoints, we shall conform to standard usage in this regard. Before doing so, however, we offer the following preliminary clarification of the approach we shall adopt.

We have seen (in Chapter 11, Sections 11.3 and 11.4) that the notions of exchangeability and partial exchangeability can be reduced to that of conditional independence (albeit sometimes in a merely formal sense). More precisely, we have seen that the probability distribution in such cases is always a mixture (i.e. a convex linear combination) of distributions representing independence. If to each case of independence there corresponds an objectively defined 'hypothesis' – like, for example, the proportion of white balls in an urn of unknown composition – then the mixture has an objectively meaningful interpretation. Where this is not the case, the representation is merely formal (as, for example, in the case of a biased coin). There is, however, no difficulty – apart from that of a conceptual nature – in dealing with such 'hypotheses' in these cases as if they were objectively meaningful: for example, one might refer, quite improperly, to 'the hypothesis that the unknown probability of obtaining Heads with the bent coin has the value p' On the other hand, this pseudo-interpretation could always be treated as an asymptotic interpretation of a property that can be defined in a finitistic way by referring to 'frequency in a large number of trials' instead of to 'unknown probability'.[1]

1 We observe that although the present approach may, at first sight, appear to be very similar (or even equivalent) to that based on 'limit-frequency', there is, in fact, a great deal of difference. We in no way assume the existence of any limit to which the frequency Y_n/n must tend (either with certainty, or in some probabilistic sense – like weak, mean-square or strong convergence and so on). Nor do we utilize any probabilistic form of Cauchy convergence (Chapter 6, 6.8.7), even though this holds (see Chapter 11, 11.4.2) under the assumption of exchangeability (it does not define a 'limit random quantity' and, in any case, requires an infinite number of 'trials'). We base ourselves solely on the frequency Y_n/n of successes in the trials actually considered, whatever they may be, and however many there are, and on the fact that their distribution F_n (the probability distribution of Y_n/n according to the evaluation made at the beginning of the trials) provides an approximation (which improves as n increases) to the limit distribution F, whose existence is therefore established (and this is the only thing we need!).

We shall therefore adopt, in line with our previous remarks, the standard practice of talking in terms of 'hypotheses', irrespective of whether these exist objectively, or merely formally (in which case, they might be interpreted, if at all, in the asymptotic sense given above). Within this framework, the Bayesian approach consists in considering an initial distribution of probability among these hypotheses, this distribution being modified as new information becomes available.

12.2.2. The enormous range of possible applications might lead one to expect a large number of different theoretical models. On the other hand, if one thinks in terms of the basic simple forms of representation, the possibilities are more limited. (Indeed, one might argue that from a Bayesian point of view there is only one form of the problem since everything reduces, in the final analysis, to an application of Bayes's theorem.) In fact, those cases which form the bulk of mathematical statistics can probably be reduced to one or other of the two forms already mentioned: exchangeability or partial exchangeability. There is a meaningful distinction to be drawn between these two cases and, indeed, partial exchangeability embraces a wide range of possible deviations from the exchangeable case. Moreover, within the two categories there are a number of problems of detail and, depending on the field of events under consideration, these may present various levels of difficulty (without there necessarily being any great conceptual difficulties).

Exchangeability and partial exchangeability can also arise in the context of multi-events in general (as we mentioned in Chapter 11), as well as for vectors (r-tuples of random quantities), functions,..., and random elements of any space whatsoever.

What is important in these cases is not so much their actual form, or that of the space to which they belong, but rather the kinds of 'hypotheses' that are assumed; that is the corresponding distributions. There are three main distinctions worth making in this respect: the *discrete* case (involving a finite or countable set of hypotheses); the *parametric* case (involving a set of hypotheses which can each be represented in terms of a fairly restricted set of parameters; i.e. by a vector in a parameter space whose dimension is not too large); *the nonparametric case* (where either the individual hypothesis cannot be represented in terms of a finite number of parameters, or, alternatively, the number of parameters involved is prohibitively large).

Similar distinctions can be made in the case of forms of representation required for *partial* exchangeability (and we shall shortly give a more precise account of these).

12.2.3. In presenting the mathematical development of these ideas, we shall normally deal with problems involving random quantities in the parametric case; in particular, with just one parameter. This is the most straightforward and meaningful case, and hence the most convenient for illustrating the mathematics. It will be immediately obvious, however, that our treatment is completely general, provided the expressions given and the comments made are interpreted in an appropriate manner.

Specifically – to use what we regard as the correct terminology – we shall be dealing with a collection of exchangeable random quantities X_h (see Chapter 11, Sections 11.3 and 11.4). These can be represented in precisely the same way as we saw earlier in the case of exchangeable events: in other words, they can, in a formal sense at least, be thought of as 'independent conditional on some given set of hypotheses'. More precisely, each 'hypothesis' indexes a distribution and we assume that, conditionally, 'all the X_h have this same distribution, and are stochastically independent of one another'.

This implies that their joint distribution is a mixture of products of individual factors (corresponding to the case of independence). As we remarked earlier on, we shall take this representation as our starting point; the interpretation in terms of exchangeability then becomes merely a preliminary clarification.

We shall follow the usual practice in mathematical statistics and work in terms of probability densities (for a justification of this, see the remarks in Section 12.4.3). Expressed mathematically, our basic assumptions then become the following:

- there exists a set of 'hypotheses', a general element of this set being denoted by θ (a point in the hypothesis space); in particular, we shall first consider the case where each hypothesis can be represented by a single real-valued parameter θ;
- conditional on each hypothesis θ (i.e. on the value of θ for the case in question), all the X_h have exactly the same distribution, that is the same density $f(x|\theta)$, and are stochastically independent; this implies that the joint density $p^m(x^1,\ldots, x^m|\theta)$[2] for m of the X_h (no matter how they are chosen or labelled) is given by the product of the densities

$$p^m\left(x^1,x^2,\ldots,x^m\,|\,\theta\right)=f\left(x^1\,|\,\theta\right)f\left(x^2\,|\,\theta\right)\ldots f\left(x^m\,|\,\theta\right);$$

- over the set of hypotheses we have prior probability with density $\pi_0(\theta)$; in the case we are considering, we have a non-negative function of θ such that

$$\int_{-\infty}^{\infty}\pi_0\left(\theta\right)\mathrm{d}\theta=1.$$

We note that this latter assumption is the hallmark of the Bayesian approach, whereas other approaches attempt to do without it. We shall develop our treatment within the Bayesian framework but, as we proceed, we shall discuss the techniques that are used by those who eschew the use of 'prior probabilities'.

It follows immediately from these assumptions that the marginal (prior) distributions for any individual X_h, or for m of them, are, expressed as densities, given by

$$f_0\left(x\right)=\int f\left(x\,|\,\theta\right)\pi_0\left(\theta\right)\mathrm{d}\theta, \tag{12.1}$$

$$\begin{aligned} p_0^m\left(x^1,x^2,\ldots,x^m\right)&=\int p^m\left(x^1,x^2,\ldots,x^m\,|\,\theta\right)\pi_0\left(\theta\right)\mathrm{d}\theta\\ &=\int f\left(x^1\,|\,\theta\right)f\left(x^2\,|\,\theta\right)\ldots f\left(x^m\,|\,\theta\right)\pi_0\left(\theta\right)\mathrm{d}\theta. \end{aligned} \tag{12.2}$$

Here, and elsewhere, it is to be understood that the integrals are to be taken over the entire range of the distribution (and there is no harm in thinking of this as the whole real line, since any range where the density is zero will give a zero contribution). Note that, if we interpret the quantities involved in an appropriate manner, these expressions apply equally well to any abstract spaces (and, in particular, to the case of several parameters, where θ represents a vector).

From the point of view of interpretation, note that f_0 and p_0^m give the previsions of f and p^m if the latter are considered as functions of the random quantity θ. Also note that

---

2 The reason for using superscripts will become clear in Section 12.2.5. Note that f is a special case of p^m for $m=1$ ($f=p^1$, and, in what follows, $f_n=p_n^1$, etc.). Usually, however, we shall use f in preference to p^1 in order to make the case $m=1$ more immediately distinguishable, and also to avoid the superscript.

p_0^m, like p^m, is a symmetric function of the x^h (in line with our original assumption of exchangeability).

12.2.4. It is equally straightforward to derive expressions, similar to the above, for the evaluations conditional on knowledge of the values of any of the $X_{h,}$ or of any n of them. We shall denote these random quantities by X_1, or X_1, X_2,..., X_n, partly for convenience and partly to fix ideas for the case in which we observe them in chronological order (and although this might be useful, the reader should remember that it is not an essential part of the argument). The choice of which particular X_h (or set of them) we are interested in calculating the conditional evaluation for is equally irrelevant and the reader should again realize that we denote these by

$$X_{n+1}, X_{n+2}, \ldots, X_{n+m}$$

purely for convenience.

We shall see, in fact, that the evaluations conditional on the knowledge of the values of the first n of the X_h (which we shall denote by f_n and p_n^m) can be expressed in essentially the same form as the f_0 and p_0^m above (the special cases corresponding to $n = 0$; i.e. the initial evaluations, prior to any knowledge of the X_h). In fact, it turns out to be sufficient to determine the distribution $\pi_n(\theta \mid x_1, x_2, \ldots, x_n)$[3] for the parameter θ, conditional on the values $X_1 = x_1$, $X_2 = x_2$,..., $X_n = x_n$, and to substitute this in place of $\pi_0(\theta)$ in the expressions for f_0 and p_0^m. Let us first see this for the case $n = 1$.

After having observed the value of any one of the X_h, $X_1 = x_1$, say, the probability distribution of the parameter (or, more accurately,[4] the distribution 'conditional on the hypothesis $X_1 = x_1$') becomes

$$\pi_1(\theta \mid x_1) = K \pi_0(\theta) f(x_1 \mid \theta) \quad \left(\frac{1}{K} = \int f(x_1 \mid \theta) \pi_0(\theta) \mathrm{d}\theta = f_0(x_1) \right). \tag{12.3}$$

This is a straightforward application of Bayes's theorem, or, if one prefers, it is sufficient to observe that the joint density for (θ, X_1) is given both by $\pi_0(\theta) f(x_1 \mid \theta)$ and by $f_0(x_1) \pi_1(\theta \mid x_1)$.

It follows immediately that

$$f_1(x \mid x_1) = \int f(x \mid \theta) \pi_1(\theta \mid x_1) \mathrm{d}\theta, \tag{12.4}$$

$$p_1^m(x^2, x^3, \ldots, x^{m+1} \mid x_1) = \int p^m(x^2, x^3, \ldots x^{m+1} \mid \theta) \pi_1(\theta \mid x_1) \mathrm{d}\theta. \tag{12.5}$$

12.2.5. Before we go on with our development, we need to make a comment about notation. The use of superscripts for the x (x^h rather than x_h) is necessary in order to distinguish the use of the values x^h of X_h as 'names of coordinates' for the distribution of X_h (as yet unknown, or considered as unknown), from the use of the x_h as observed values (or values

3 N.B. For the sake of brevity, we shall sometimes write this as $\pi_n(\theta)$ omitting any explicit mention of $x_1, x_2, \ldots, x_n$ (which must of course be understood).

4 More accurately' by virtue of what we said in Chapter 11, 11.2.2 (and what we shall say in Section 6 of the Appendix).

assumed to be known). The practical effect of this can be observed in the formulae, where superscripts precede the vertical bar and subscripts follow it (except when rôles are reversed, as happens during the application of Bayes's theorem; see equations 12.11 and 12.11)). In the case of something like $f_1(x|x_1)$, it is clear that it would be superfluous to write $f_1(x^2|x_1)$, because the superscripts are only useful for distinguishing between the x^h when there is more than one of them. For a single (generic) coordinate, it is sufficient to denote it by x.

12.2.6. The expressions for f_1 and p_1^m (and we recall that the former is a special case of the latter; $f_1 = p_1^1$) can be rewritten in a different form, so as to show up certain interesting features more clearly:

$$f_1(x\,|\,x_1) = K\int f(x\,|\,\theta)\big[f(x_1\,|\,\theta)\pi_0(\theta)\mathrm{d}\theta\big] \qquad \left(\frac{1}{K} = \int[\ldots] = f_0(x_1)\right); \tag{12.6}$$

$$\begin{aligned} p_1^m\left(x^2,x^3,\ldots,x^{m+1}\,|\,x_1\right) &= K\int p^m\left(x^2,x^3,\ldots,x^{m+1}\,|\,\theta\right)\big[f(x_1\,|\,\theta)\pi_0(\theta)\mathrm{d}\theta\big] \\ &= K\int\prod_{i=2}^{m+1} f\left(x^i\,|\,\theta\right)\big[f(x_1\,|\,\theta)\pi_0(\theta)\mathrm{d}\theta\big]. \end{aligned} \tag{12.7}$$

In this way, we emphasize the fact that f_1 and p_1^m are mixtures of f and p^m, with 'weights' as given in square brackets. Alternatively, we could remove the separation between factors in x_1 and those in x^i ($i = 2, 3,\ldots, m+1$) and write instead

$$\begin{aligned} f(x\,|\,x_1) &= K\int\left\{f(x\,|\,\theta)f(x_1\,|\,\theta)\right\}\pi_0(\theta)\mathrm{d}\theta \\ &= Kp_0^2(x_1,x) = \frac{p^2(x_1,x)}{f_0(x_1)}, \end{aligned} \tag{12.8}$$

or even

$$f_1(x\,|\,x_1) = f_0(x)\frac{f_1(x_1\,|\,x)}{f_0(x_1)}. \tag{12.9}$$

In this way, we directly emphasize the interpretation in terms of the theorem of compound probabilities and Bayes's theorem.

Proceeding in a similar fashion, we can derive the more general result

$$p_1^m\left(x^2,x^3,\ldots,x^{m+1}\,|\,x_1\right) = \frac{p_0^{m+1}\left(x_1,x^2,x^3x,\ldots,x^{m+1}\right)}{f_0(x_1)}. \tag{12.10}$$

In order to derive a form analogous to that of equation 12.9, that is

$$p_1^m\left(x^2,x^3,\ldots,x^{m+1}\,|\,x_1\right) = \frac{p_0^m\left(x^2,x^3,\ldots,x^{m+1}\right)f_m\left(x_1\,|\,x^2,x^3,\ldots,x^{m+1}\right)}{f_0(x_1)}, \tag{12.11}$$

we must introduce the f_m for $m > 1$; the result is then immediate.

12.2.7. It would have been perfectly straightforward, and more in line with the approaches more commonly adopted in statistics, to have considered right away the distributions conditional on n values,

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$$

instead of on just one. Our main consideration in starting off with the case $n = 1$ is that it enables one to bring out the fact that the effect of n observations is simply the combined effect of considering them one at a time, rather than some magical consequence of there being enough of them to be 'statistically' relevant.

The probability distribution of the parameter θ, given the observed values $x_1, x_2, \ldots, x_n$, is given by

$$\begin{aligned}\pi_n(\theta \mid x_1, x_2, \ldots, x_n) &= K\pi_0(\theta)\, p^n(x_1, x_2, \ldots, x_n \mid \theta) \\ &= K\pi_0(\theta) f(x_1 \mid \theta) f(x_2 \mid \theta) \ldots f(x_n \mid \theta),\end{aligned} \tag{12.3′}$$

where

$$\frac{1}{K} = \int p^n(x_1, x_2, \ldots, x_n \mid \theta)\pi_0(\theta)\,\mathrm{d}\theta = p_0^n(x_1, x_2, \ldots, x_n).$$

As we remarked earlier, it is sufficient to replace π_0 by π_n in order to obtain the distributions for one of the X_h, or for m of them:

$$f_n(x \mid x_1, x_2, \ldots, x_n) = \int f(x \mid \theta)\pi_n(\theta \mid x_1, x_2, \ldots, x_n)\,\mathrm{d}\theta \tag{12.4′}$$

$$= K\int f(x \mid \theta)\left[\prod_{h=1}^{n} f(x_h \mid \theta).\pi_0(\theta)\,\mathrm{d}\theta\right] \tag{12.6′}$$

$$= p_0^{n+1}(x_1, x_2, \ldots, x_n, x) / p_0^n(x_1, x_2, \ldots, x_n) \tag{12.8′}$$

$$= f_0(x).p_1^n(x_1, x_2, \ldots, x_n \mid x) / p_0^n(x_1, x_2, \ldots, x_n); \tag{12.9′}$$

$$\begin{aligned}&p_n^m(x^{n+1}, x^{n+2}, \ldots, x^{n+m} \mid x_1, x_2, \ldots, x_n) \\ &\quad = \int p^m(x^{n+1}, \ldots, x^{n+m} \mid \theta)\pi_n(\theta \mid x_1, \ldots, x_n)\,\mathrm{d}\theta\end{aligned} \tag{12.5′}$$

$$= K\int \prod_{i=n+1}^{n+m} f(x^i \mid \theta).\prod_{h=1}^{n} f(x_h \mid \theta)\pi_0(\theta)\,\mathrm{d}\theta \tag{12.7′}$$

$$= p_0^{n+m}(x_1, \ldots, x_n, x^{n+1}, \ldots, x^{n+m}) / p_0^n(x_1, \ldots, x_n) \tag{12.10′}$$

$$= \frac{p_0^m(x^{n+1}, \ldots, x^{n+m})\, p_m^n(x_1, \ldots, x_n \mid x^{n+1}, \ldots, x^{n+m})}{p_0^n(x_1, \ldots, x_n)}. \tag{12.11′}$$

The last four expressions include all the others as special cases: more precisely,

equations 12.5′, 12.7′, 12.10′ and 12.11′ give for general n and m, what equations 12.5, 12.7, 12.10 and 12.11 give for $n = 1$ (with m arbitrary), and

equations 12.4′, 12.6′, 12.8′ and 12.9′ give for $m = 1$ (with n arbitrary) and equations 12.4, 12.6, 12.8 and 12.9 give for $n = m = 1$.

The interpretations are identical to those that we gave in the simplest case (i.e. that of $n = m = 1$) and, when contemplating extensions to cases that are more complicated (insofar as the formulae are concerned, anyway), it may be useful to bear this case in mind.

Given $x_1, x_2, \ldots, x_n$, the likelihoods for θ and x are

a) $\prod_h f(x_h \mid \theta)$ (as a function of θ),

b) $\int f(x \mid \theta) \prod_h f(x_h \mid \theta) \pi_0(\theta) \mathrm{d}\theta$ (as a function of x),

respectively. In fact, any function differing from (a) by a factor independent of θ, or from (b) by a factor independent of x, could be taken as the respective likelihood.

12.2.8. We now turn to the case of 'partial exchangeability'. The account we shall give will be even shorter than the above and we shall rely on the examples to clarify our interpretation and approach to the problem.

In terms of our formulation, this case differs from the previous one in that, conditional on each of the 'hypotheses' characterized by θ, the X_h are still stochastically independent, but now may have different distributions. The latter depend not only on the parameter θ, but also on certain observable quantities y_h, which relate to the X_h. Like θ, y may be real-valued, or a vector, or whatever (irrespective of the form of θ).[5] In order to keep the presentation on a simple level, we shall take y to be real-valued (the general case presents nothing new from a conceptual viewpoint).

Formally, instead of starting from $f(x|\theta)$ we consider $f(x|\theta, y)$. So far as the prior distribution for θ is concerned, nothing changes; we begin, as before, with some $\pi_0(\theta)$. The distribution $p_0^m(x^1, x^2, \ldots, x^m)$ (knowledge of which enables one to derive everything else) will, however, also depend on the values that y takes for each of the $X_x, X_2, \ldots, X_m$, and has the form

$$p_0^m\left(x^1, x^2, \ldots, x^m\right) = \int f\left(x^1 \mid \theta, y_1\right) f\left(x^2 \mid \theta, y_2\right) \ldots f\left(x^m \mid \theta, y_m\right) \pi_0(\theta) \mathrm{d}\theta, \quad (12.12)$$

where y_h corresponds to X_h. If we wish this to be made explicit, we must write the left-hand side as

$$p_0^m\left(x^1, \ldots, x^m \mid y_1, \ldots, y_m\right).$$

On the other hand, if we do make systematic use of this explicit form the expressions become rather cumbersome – particularly those which are already complicated, even without this additional detail.

The following are intended as examples of the kinds of y_h that might be observed[6] and considered as possibly influencing the distribution of X_h: the temperature at the time at

5 In other words, one could be a vector and the other a real number, etc.

6 See the remark at the end of Section 12.3.3.

which the experiment yielding X_h took place; the age of an individual whose reaction to some given drug is measured by X_h; the precision of the instrument which performs the measurement giving X_h; and so on. We shall shortly give an example involving the latter possibility.

12.3 Examples Involving the Normal Distribution

12.3.1. Given that the normal distribution is widely used (and somewhat abused) in statistics, it is natural that the most familiar problems of inference are those which involve this distribution. The prevision m and the variance σ^2 suffice to characterize the distribution, which is usually denoted by $N(m, \sigma^2)$. The density, as we already know, is given by

$$f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{1}{2}(x-m)^2/\sigma^2\right\}.$$

By far the most important case is that in which m corresponds to the unknown parameter θ (while σ^2 is known), but we shall also deal with the opposite case ($\theta = 1/\sigma^2$, m known) and with the case in which both parameters are unknown (θ is the 'vector' (θ_1, θ_2), $\theta_1 = m$ and $\theta_2 = 1/\sigma^2$).[7] On the basis of these examples, all under the assumption of complete exchangeability, we can discuss variants corresponding to partial exchangeability. The simplest such variants are obtained by replacing the assumption 'σ^2 known' (or 'm known') by '$= y$', known for each X_h, but possibly different for different h'.

12.3.2. *The case where m is unknown*. This is, above all, the case considered in the theory of errors (experimental or observational) as applied in astronomy geodesy, physics and so on. What is unknown is the *true* value of the quantity that is being measured; that is $\theta = m$. The accuracy of the instrument (as represented by σ^2) is assumed known and the distribution of the observed value is assumed to be $N(m, \sigma^2)$; that is a normal distribution centred at the *true* value and having the given precision.

We have, therefore,

$$f(x \mid \theta) = K \exp\left\{-\frac{1}{2}(x-\theta)^2/\sigma^2\right\}, \tag{12.13}$$

and (apart from the constant factor K) this is the likelihood for θ given by an observation x. The likelihood given by n observations $x_1, x_2, \ldots, x_n$ is

$$\prod_h f(x_h \mid \theta) = \exp\left\{-\frac{1}{2\sigma^2}\sum_h (x_h - \theta)^2\right\}.$$

---

7 In the terminology used in the theory of errors, the reciprocal $1/\sigma$ of the standard deviation is called the *precision*, and the reciprocal $1/\sigma$ of the variance is called the *weight* (although sometimes a different unit of measure is used; e.g. precision $1/\sqrt{2}\sigma$, weight σ_0^2/σ^2, where σ_0^2 is chosen as appropriate for the problem under consideration). It might seem rather unnecessary to have four terms available, but this is not entirely the case.

We shall take the *weight* σ^{-2} as the parameter θ, instead of the more customary variance, σ^2, since this turns out to simplify the formulae.

Noting that

$$\sum_h (x_h - \theta)^2 = \sum_h \left(x_h^2 - 2x_h\theta + \theta^2\right) = \text{const.} - 2\theta\sum_h x_h + n\theta^2$$
$$= n\left(\text{const.} - 2\theta\frac{1}{n}\sum_h x_h + \theta^2\right)$$
$$= n\left[\text{const.} + (\bar{x} - \theta)^2\right],$$

where $\bar{x} = 1/n\sum_h x_h$ the mean of the x_h, we see that the likelihood can be rewritten in the form

$$\exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \theta)^2\right\}. \tag{12.14}$$

In other words, it has the same form as the likelihood of a single observation equal to the mean $\bar{x}$, and with *standard deviation* $\sigma/\sqrt{n}$ (i.e. reduced in the ratio 1 to $1/\sqrt{n}$, which is equivalent to *precision* increased in the ratio 1 to $\sqrt{n}$, *variance* reduced in the ratio 1 to $1/n$, *weight* increased in the ratio 1 to n).

The posterior distribution for θ is therefore given by

$$\pi_n(\theta) = K\pi_0(\theta)\exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \theta)^2\right\}. \tag{12.15}$$

Since the likelihood is maximized for $\theta = \bar{x}$ and decreases as we move away from this value (the decrease being sharper for larger n), the posterior distribution concentrates around x. In particular, if the prior distribution is taken to be normal, $N(m_0, \sigma_0^2)$, say then the posterior distribution is also normal. More precisely, the posterior (or final[8]) distribution is $N(m_\text{f}, \sigma_\text{f}^2)$, where

$$m_\text{f} = \frac{m_0\sigma_0^{-2} + n\bar{x}\sigma^{-2}}{\sigma_0^{-2} + \sigma^{-2}}, \qquad \sigma_\text{f}^{-2} = \sigma_0^{-2} + n\sigma^{-2}. \tag{12.16}$$

In words: the posterior *weight* (1/*variance*) is the sum of the weights from the prior and the likelihood; the posterior mean is the weighted mean of the prior mean and the mean from the likelihood (m_0, and n times $\bar{x}$; i.e. a function of m_0 and $x_1, x_2, \ldots, x_n$), the weights being the respective *weights* (thus revealing the aptness of the terminology).

12.3.3. The extension to the case of 'observations made with different precisions' is immediate. Let us assume, for instance, that we know that the n observations are performed with different measuring instruments, the errors of which have standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$. It is clear that (by an argument similar to that used above) these observations are equivalent to a single observation whose value is given by the weighted mean of the x_h (with weights σ_h^{-2}) and having weight equal to the sum of the weights.

---

8 *Translators' note.* The terms *prior* and *posterior* seem firmly established in English publications relating to applications of Bayes's theorem, and we have used them in preference to the terms *initial* and *final*. The Italian version uses the latter, and the notation m_f and σ_f^2 derives from this usage.

If the prior distribution is $N(m_0, \sigma_0^2)$, the posterior distribution is given by $N(m_f, \sigma_f^2)$, where m_f and σ_f^2 are determined by the weighting process just described, except that we now also include m_0 with weight σ_0^{-2}.

This is an example of 'partial exchangeability' with $y_h = \sigma_h^2$ (or we could take $y_h = \sigma_h^{-2}$, and

$$f(x \mid \theta, y) = K \exp\left\{-\frac{1}{2}(x-\theta)^2 / y\right\}.$$

We should draw attention to the fact that the y_h must actually be known and observed for each X_h under consideration. In our example, we must know with what precision each measurement has been performed. One should be careful not to think of it as being sufficient to know that each measurement has been performed using instruments of various precisions (for example, by choosing each time at random from among some given collection of measuring instruments, but without registering which were actually used and how often). Under this latter assumption, one would have a case of exchangeability with

$$f(x \mid \theta) = \sum_k c_k f_k(x \mid \theta), \quad f_k(x \mid \theta) = K \exp\left\{-\frac{1}{2}(x-\theta)^2 y_k\right\}, \quad c_k \geqslant 0, \quad \sum_k c_k = 1$$

(i.e. no longer a normal distribution but a mixture of normals). In the same way, in the other examples it would be necessary to have actually noted the temperature, age etc., in each case.

12.3.4. *Comments*. The choice of the normal form for the prior distribution in the case just considered is convenient in that the posterior distribution is then always a member of this same family. We shall see that in other cases, too, we can find distributions for which this property holds.

On the other hand, this convenience does not justify our making such a choice if it is not compatible with our actual prior opinion; neither does it provide any *a priori* justification for regarding such distributions as in any way playing a special role. A reasonable approach involves adopting 'convenient' distributions if and insofar as they provide a sufficiently accurate representation of one's actual opinions (and this is especially useful in those problems where the precise form chosen has little influence on the final outcome).

If the influence on the final outcome is going to be practically negligible, one might even 'omit' the factor $\pi_0(\theta)$ altogether; that is, to be more precise, one might consider the limit case of a 'constant density'. This improper distribution could be interpreted, for example, as the limit of the normal distribution $N(0, \sigma^2)$ as $\sigma^2 \to \infty$, or of the uniform distribution

$$\pi_0(\theta) = \frac{1}{2}a\left(-a \leqslant \theta \leqslant a\right) \text{ as } a \to \infty.$$

As we shall see, other forms of improper prior distribution may be more appropriate, depending on the form of the problem.

Sometimes the use of the improper, uniform prior distribution is interpreted as representing 'total ignorance'. This is nonsense: every distribution reflects some sort of

opinion, and none of these have any special status – not even in the negative sense of representing no opinion at all. Moreover, one should note that the uniform distribution is not invariant under changes in the parametrization (e.g. θ into log θ or e^{θ} etc.). A number of useful observations of this kind can be found in Lindley, Vol. II (with particular reference to this topic, see p. 145).

Remarks. The above considerations are all dependent on a certain mathematical point which should be clearly understood, because it serves to clarify the particular practical consequences of the above.

Rigorously speaking, a *density is not just a point function but rather a function of the point and of the measure* assumed over the space under consideration. For instance, it is well known that in the case of measures defined in terms of coordinate systems a change of coordinates alters the density by multiplying it by the Jacobian (and the same thing holds more generally). It follows, for instance, that we could always arrange to have a constant density (it suffices to take the distribution corresponding to such a density as the underlying measure).

The *likelihood*, on the other hand, actually is a point function, and 'equating' it to a density is a meaningless idea. We can always achieve what we want, however, by an appropriate choice of the measure (which is never significant from a theoretical viewpoint), taking it over the most convenient reference system (or one which is sufficiently convenient) in order to make calculations as straightforward as possible.

We shall, therefore, find it useful (and a number of examples of this will be given) to choose a family of prior distributions with density 'equal' to the likelihood. In the terminology introduced by Raiffa and Schlaifer, these constitute the *conjugate* family for the problem. One should note, however, that this notion has no absolute meaning, but can be useful relative to some given standard formulation of a problem.

12.3.5. *The case where σ^2 is unknown.* We again consider the normal distribution, $N(m, \sigma^2)$ but now with the variance σ^2 unknown (and it is convenient to set $\theta = 1/\sigma^2$ = 'weight') and the mean m known. This case arises, for example, if one wishes to calibrate a new measuring instrument (i.e. to determine its precision as measured by $\theta = 1/\sigma^2$) by making repeated measurements of a given known quantity m.

In this case we have

$$f(x\,|\,\theta) = K\theta^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\theta(x-m)^2\right\}. \tag{12.17}$$

As a function of θ (and leaving aside factors independent of θ), this is the likelihood for θ given by an observation x.

The likelihood for θ given by n observations $x_1, x_2, \ldots, x_n$ is, therefore,

$$\prod_h f(x_h\,|\,\theta) = \theta^{\frac{1}{2}n} \exp\left\{-\frac{1}{2}\theta\sum_h (x_h-m)^2\right\} = \theta^{\frac{1}{2}n} e^{-\frac{1}{2}\theta S^2}, \tag{12.18}$$

where $S^2 = \sum_h (x_h - m)^2$ (the constant K^n having been omitted).

Since the form of this expression is that of the density of a gamma distribution, any choice of prior from within the gamma family will ensure that the posterior distribution

belongs to the same family (and the comments of Section 12.3.4 should be understood in this case, too). Taking

$$\pi_0(\theta) = K\theta^{\alpha-1}\mathrm{e}^{-\lambda\theta},$$

we obtain

$$\pi_0(\theta) = K\theta^{\alpha-1+\frac{1}{2}n}\mathrm{e}^{-\left(\lambda+\frac{1}{2}S^2\right)\theta}. \tag{12.19}$$

12.3.6. *The case where both m and σ^2 are unknown*. This arises in the context of errors of observation, as in Section 12.3.2, except that we also assume the precision of the measuring instrument to be unknown. It is also the most frequently studied case in statistics, where we have a population (of individuals, objects, experiments etc.) in which some given quantity (X_h for the *hth* individual) is known (or assumed) to be normally distributed, but with neither the mean (central) value nor the variance known.

The example involves two parameters – those encountered separately in the previous two cases. We put $\theta_1 = m$, $\theta_2 = 1/\sigma^2$, and hence we obtain

$$f(x\,|\,\theta_1,\theta_2) = K\theta_2^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\theta_2(x-\theta_1)^2\right\}. \tag{12.20}$$

The likelihood for θ_1 and θ_2 after having observed $x_1, x_2,\ldots, x_n$ is given by

$$\theta_2^{n/2}\exp\left\{-\frac{1}{2}\theta_2\sum_h(x_h-\theta_1)^2\right\} = \theta_2^{n/2}\exp\left\{-\frac{1}{2}\theta_2\left[\nu s^2 + n(\bar{x}-\theta_1)^2\right]\right\}, \tag{12.21}$$

where

$$\nu = n-1, \quad s^2 = \sum_h(x_h-\bar{x})^2/\nu, \quad \bar{x} = \sum_h x_h/n$$

(the steps are the same as those given in Section 12.3.2, except that the constant νs^2 can now no longer be omitted because it is multiplied by the parameter θ_2).

The standard assumption for the prior distribution is, in this case, the improper uniform distribution for both θ_1 and $\log\theta_2$: this results in an improper 'density proportional to $1/\theta_2$'

$$\left(\text{over the half-plane} -\infty < \theta_1 < +\infty, 0 < \theta_2 < +\infty\right).$$

Strictly speaking, this assumption is only made by Bayesians (and even then only by those who have no objections to improper distributions), but there is some justification for referring to it as the standard assumption because it leads to the same conclusions as those arrived at by non-Bayesian statisticians using other methods of approach.

With these assumptions, that is supposing that we are prepared to express the prior density in the form

$$\pi_0(\theta_1,\theta_2) = K/\theta_2, \tag{12.22}$$

we obtain

$$\pi_n\left(\theta_1,\theta_2\right)=K\theta_2^{(n-2)/2}\exp\left\{-\frac{1}{2}\theta_2\left[\nu s^2+n\left(\bar{x}-\theta_1\right)^2\right]\right\}. \tag{12.23}$$

The marginal posterior densities for θ_1 and θ_2 (obtained by integrating out the other variable) are[9]

$$\pi_n^{(1)}\left(\theta_1\right)=K\left\{1+n\left(\bar{x}-\theta_1\right)^2/\nu s^2\right\}^{\frac{1}{2}n}=K\left(1+t^2/\nu\right)^{-\frac{1}{2}(\nu+1)} \qquad \left(t=\sqrt{n}\frac{\bar{x}-\theta_1}{s}\right), \tag{12.24}$$

$$\pi_n^{(2)}\left(\theta_2\right)=K\theta_2^{\nu/2}\exp\left\{-\frac{1}{2}\nu s^2\theta_2\right\}. \tag{12.25}$$

For θ_2, we still have a gamma distribution, just as in the case m known (Section 12.3.5). For θ_1, on the other hand, the normal distribution, which we obtained in the case σ^2 known (Section 12.3.2), has been replaced by Student's distribution (see the very end of Chapter 10). The effect of not knowing m and σ^2 is that they are replaced by $\bar{x}$ and s^2 (which are 'reasonable estimates' of them), and that, in the case where we are ignorant of σ^2, the normal is replaced by the Student distribution which has much fatter tails (although it tends to the normal as $n\to\infty$). For large n, therefore, the difference is practically negligible.

12.4 The Likelihood Principle and Sufficient Statistics

12.4.1. Given that we started out by adopting the Bayesian approach and that we have adhered to it coherently throughout, the 'likelihood principle' inevitably appears to be rather obvious and certainly not worth getting excited about. It simply states that the information available from any set of observations is entirely contained in the corresponding likelihood function. Since this is, in fact, the factor which transforms the prior opinion into the posterior, this is all we require and, indeed, all we can ask for.

If, however, one proposes ad hoc methods – more or less on a trial and error basis – it might well happen that they conflict with this 'principle'. For this reason, non-Bayesians have debated among themselves as to whether this principle (or a variant thereof) should be rejected, or whether, on the contrary, one should reject methods which do not comply with it (or whether such methods could be considered valid as approximations).

From the Bayesian standpoint there are two possible reasons for wishing to mention the principle: firstly, to warn against superficial interpretations of it; secondly, as a starting point for developing the topic of 'sufficient statistics'.

---

9 For the details of the calculations, see, for example, Lindley, 5.3 and 5.4, but note that he takes $\theta_2=\sigma^2$ (whereas by setting $\theta_2=1/\sigma^2$, we have obtained a certain amount of simplification in the formulae and calculations).

The warnings are the obvious ones (but, on the other hand, mistakes are often the result of overlooking the obvious) and concern a too literal interpretation of the following statement of the principle:

> *The information contained in a set of observations is completely summarized by the likelihood function, and provided it can be combined with similar results etc., it is quite sufficient to quote this.*

All is well, provided we also enter the following reservation: '*so long as the basic assumptions remain unchanged*'. If, for instance, one starts off by assuming that certain errors 'are normally distributed' and then begins to wonder whether they, in fact, have some other distribution, the data expressed in the form of the 'likelihood function' can no longer provide all relevant information.

The discussion about '*sufficient statistics*' could begin by our pointing out that the likelihood function is itself a sufficient statistic (that is to say, it provides an exhaustive summary of the information contained in the data). It follows, therefore, that the summaries of the data that characterize the likelihood, when considered altogether, themselves form a sufficient statistic. In the cases that we have examined, for example, we have the following:

Section	Case	Sufficient statistic	
12.3.2	(m unknown, σ^2 known and constant)	the pair	$n,\ \bar{x}$;
12.3.3	(m unknown, σ^2 known but varying)	the pair	$\sum y_h, \sum y_h x_h$;
12.3.5	(σ^2 unknown, m known and constant)	the pair	n, S^2;
12.3.6	(m and σ^2 both unknown)	the triple	$n,\ \bar{x},\ s^2$.

12.4.2. For the sake of completeness, we now give a few of the basic notions relating to the concept of a sufficient statistic.

In what follows, it will be convenient to denote the *data* (which in general consists of n observations, x_h; $h = 1, 2,\ldots, n$) simply by x; the *parameters* (no matter whether there is just one, θ, or several, θ_i, $i = 1, 2,\ldots, s$) by θ; and the *sufficient statistic* (consisting either of a single real-valued function of the data, $t = t(x)$, or of several; $t_j = t_j(x)$; $j = 1, 2,\ldots, r$) by t; in the same way $p(x|\theta)$ denotes something of the form $p^n(x_1, x_2,\ldots, x_n|\theta)$ and so on. From a conceptual point of view, the argument is precisely the same, whether $x, \theta, t,\ldots$ are real numbers, or vectors, or whatever.

Expressing our comments about sufficient statistics in the form of a definition, we have the following (also known as the 'sufficiency principle'): $t(x)$ is a sufficient statistic for the family $p(x|\theta)$ if and only if, for any prior $\pi_0(\theta)$, the posterior distribution is the same no matter whether we condition on x or on $t(x)$; that is. $\pi(\theta|x) = \pi(\theta|t(x))$.

A necessary and sufficient condition for $t(x)$ to be a sufficient statistic for $p(x|\theta)$ is that the latter can be written as

$$p(x\,|\,\theta) = f\big(t(x),\theta\big).g(x), \tag{12.26}$$

where f and g are arbitrary functions. The necessity of the condition is obvious. From the definition, it follows that

$$p(x|\theta) = p\big(t(x)|\theta\big)\,p\big(x|t(x),\theta\big),$$

and this is then equal to

$$p\big(t(x)|\theta\big)\,p\big(x|t(x)\big).$$

If we now take the first factor as f and the second as g this is in the required form. The sufficiency part is known as *Neyman's factorization theorem*.

What we have stated above is true in the general case (i.e. x does not necessarily have to consist of the results of 'independent observations from the same distribution'). If we go back to the special case (complete exchangeability), we can pose some further problems. In this case, we may be interested in knowing, for example, whether, as n varies, we can always obtain a sufficient statistic of fixed dimension (for example, having r components: $t = (t_1(x), t_2(x),\ldots,t_r(x))$).

Ignoring the finer points, the condition for this to happen is that the family of distributions $f(x|\theta)$ is a member of the *exponential family*: that is that it is of the form

$$f(x|\theta) = F(x)G(\theta).\exp\left\{\sum_{j=1}^{r} u_j(x)\Phi_j(\theta)\right\}, \tag{12.27}$$

where F, G, u_j, Φ_j, are arbitrary functions. In this case, a sufficient statistic, given any n observations $x = (x_1, x_2,\ldots, x_n)$, is provided by the r functions

$$t_j(x) = \sum_{i=1}^{n} u_j(x_i) \qquad (j = 1,2,\ldots,r), \tag{12.28}$$

together with n (although the latter is sometimes left to be understood).
In this case, the likelihood function (for θ, on the basis of the given x) is

$$p(x|\theta) = K\ .\ G(\theta)^n \exp\left\{\sum_{j=1}^{r} t_j(x)\Phi_j(\theta)\right\}. \tag{12.29}$$

If the prior distribution $\pi_0(\theta)$ is proportional to the form

$$G(\theta)^a.\exp\left\{\sum_{j=1}^{r} b_j f_{\ j}(\theta)\right\},$$

then the posterior distribution will also have this same form:

$$\pi(\theta|x) = K\pi_0(\theta)\,p(x|\theta) = KssG(\theta)^{a+n}.\exp\left\{\sum_{j=1}^{r}\big[b_j + t_j(x)\big]\Phi_j(\theta)\right\}. \tag{12.30}$$

This explains how it is always possible to define *conjugate families* of distributions whenever we are dealing with a member of the exponential family (and the same advantages are obtained as we saw previously for the normal and gamma distributions).

Recall, however, that the concept lacks any genuine substantial foundation, as we explained in the final paragraph of Section 12.3.4.

12.4.3. The time has now come for us to explain – in line with what we said in Section 12.2.3 – why we have restricted ourselves to cases in which a probability density exists.

Firstly, of course, it is quite natural to restrict oneself to the most straightforward and meaningful practical cases; these are the ones we have mentioned: either the discrete cases or those where a density exists.[10]

Over and above this, however, it is necessary to point out a far more essential reason; one which I do not think I have heard put forward before, nor have had occasion to mention myself. In order for inferences to be valid independently of the indeterminism that arises on account of 'probabilities conditional on events of zero probability', together with related questions concerning 'nonconglomerability' (Chapter 4, Sections 4.18 and 4.19, and Chapter 6, 6.9.5), it is necessary to confine oneself to problems that can be dealt with by using only probabilities conditional on hypotheses having nonzero probability. This happened trivially in the case of 'concentrated masses' and it happens directly in cases where a density exists, provided one assumes – as, fortunately, seems to be 'inevitable' from an empirical point of view – that knowledge of observed values x_h is not 'exact', but that, at best, it involves 'belonging to a neighbourhood of $(x_1, x_2, \ldots, x_n)$' small enough to make it possible to argue in terms of a density, but not in terms of the point itself.

12.5 A Bayesian Approach to 'Estimation' and 'Hypothesis Testing'

12.5.1. The natural way to present the solution of any problem of statistical inference is to give the relevant probability or probability distribution. In the cases we have considered, this involved the posterior distribution given the observed data. Unfortunately, however, such a solution cannot be regarded as 'natural', insofar as it is not 'familiar' to most people. It is for this reason, perhaps, that attempts have been made to replace the posterior distribution with some sort of crude summary conveying a more immediate message.

Two such crude approaches to summarizing the distribution of a random quantity X are widely used: the first consists in providing a unique value $\hat{x}$, around which the distribution is concentrated; the second in providing an interval $[x', x'']$, enclosing a large proportion of the distribution. These descriptions are rather vague but they can be made more precise in various ways and, in so doing, we obtain the various methods of *estimation*. More specifically, in the first case we refer to $\hat{x}$ as a *point estimate*, while in

10 Or even mixed cases; a distribution admitting a density, plus a few 'concentrated masses' at particular values of interest: for example, the percentage of some given compound in an alloy when the value zero (the absence of the compound) can occur with non-zero probability.

A typical example is the problem of King Hiero's crown (to which the episode of Archimedes' 'Eureka!' refers). Was there silver in the crown? (See L.J. Savage *et al., The Foundations of Statistical Inference*, London, Methuen (1961).) Another example is given by the correlation between two genes (0 corresponds to their being on separate chromosomes).

the second we call $[x', x'']$ an *interval estimate*. Similar considerations apply in higher dimensions (where the form of the 'interval' may be much more general).

There are other cases in which one poses the inferential question in a different way, but where one requires solutions formally similar to those given above. It may be that there is a value x_*, or an interval $[x'_*, x''_*]$, for which we wish to know whether or not X is equal to x_* (either exactly or approximately), or whether or not it lies between x'_* and x''_*. In such cases, one refers to *tests of hypotheses*, because an answer of either YES or NO is required in relation to the so-called '*null hypothesis*'

$$\left(X = x_*, \text{ or } x'_* \leqslant X \leqslant x''_*\right).$$

The contrary hypothesis, or the various hypotheses into which the complement may be divided, are known as '*alternative hypotheses*'.

12.5.2. The traditional approach to these problems, and still the most popular, is based on ad hoc methods, which, in contrast to Bayesian methods (based on a systematic and coherent theory), are largely rule-of-thumb.

In the present context, we wish to examine the extent to which they can be modified to fit into the Bayesian framework. In other words, we shall consider them not as separate and distinct methods leading to an alternative set of techniques but rather as useful summaries of certain aspects of the actual, complete solution – that is the description of the posterior distribution.

In certain respects, it is clear that the solution will depend on some value relating to the (posterior) distribution; for example, the prevision (for a fair bet), or some other mean, or the median and so on. Such a value might well be referred to as the (point) estimate for the problem; that is the appropriate *mean* in the Chisini sense.

In many cases, it is clear that giving an interval in which the random quantity of interest might plausibly be thought to lie is more informative than any attempt at actually pin-pointing it. From the Bayesian standpoint, we would give an interval having some stated probability of containing X (usually a high probability; e.g. 95%, 99%: in general, $100\beta\%$). In such a case, following Lindley, we could refer to this interval $[x', x'']$ as a $100\beta\%$ (*Bayesian*) *confidence interval for X*. The qualification 'Bayesian' will be implicit in what follows and the reader should note that a '$100\beta\%$ confidence interval' is a very different concept in a non-Bayesian context (as we shall see), and that it is important to distinguish between the two.

In general, there are infinitely many such intervals for any given level. The standard procedure is to choose the shortest one (in a certain sense, it is the most informative). In many cases – for instance, those for which the density has a unique maximum and decreases on either side of it – this interval is characterized by the fact that at each point inside it the density is greater than at every point outside it. One should note, however (in order that this criterion should not appear more 'natural' than it actually is), that both the length and the density change, in general, if X is transformed into some function of itself. For example, if $[x', x'']$ is a 95% confidence interval for X, the interval $[e^{x'}, e^{x''}]$ remains such for e^X, but if the former is the interval of shortest length, the latter, in general, is not.[11]

11 This is an obvious consequence of what we have seen more generally (see *Remark*, Section 12.3.4).

12.6 Other Approaches to 'Estimation' and 'Hypothesis Testing'

12.6.1. Those who reject the Bayesian approach cannot base their inferences on the posterior distribution even if they wished to – it does not make any sense so far as they are concerned. As a result, they are forced to have recourse to ad hoc criteria and, hence, to open the floodgates to arbitrariness. This has led to an enormous proliferation of such techniques. For the sake of completeness, and to provide a basis for certain critical comparisons, we shall give a short account of the most important and best known of these.

The basic reason why non-Bayesians are unable to refer to the posterior distribution lies in their rejection of the use of a prior distribution.[12] The best they can then do is to base themselves on the likelihood function; failing that, they simply resort to playing with formulae that are without any real foundation.

The situation can be summarized as follows.

A method for obtaining a point estimate $\hat{x}$ given the data x_h ($h = 1, 2, \ldots, n$), reduces in the final analysis, to providing a formula which expresses $\hat{x}$ as a function of the x_h: $\hat{x} = \phi_n(x_1, x_2, \ldots, x_n)$. (The same thing applies to finding the end-points x' and x'' for an interval estimate.) At the very beginning the choice of the criterion consists in defining a random quantity

$$\hat{X} = \phi_n\left(X_1, X_2, \ldots, X_n\right),$$

a function of the $X_{h,}$ whose value $\hat{X} = \hat{x}$ is to be taken as the estimate.

The problem must always be interpreted as follows (and we express it in a form which should be sufficiently vague to be acceptable to everyone): the X_h are either approximate measurements of some 'true value' x_0, which we would like to know, or they are the values of some given quantity as observed in a sample, and the value x_0 which we wish to know is some typical value (the mean, median, mode …) of the distribution of that quantity in the population. We seek to 'estimate' x_0 by $\hat{x}$.

12.6.2. There are, essentially, three different levels at which this problem can be formulated and dealt with.

At the very lowest level one simply ignores the probabilistic nature of the problem (or, at least, it is not taken into account in the formulation). At this level, we can only examine the formal properties of the proposed function and judge on empirical grounds the extent to which these are appropriate. It is rare to find this approach adhered to in any systematic way but considerations of this kind do crop up incidentally now and again

12 The paper by B. de Finetti and L.J. Savage, 'Sul modo di scegliere le probabilità iniziali', which we have already quoted several times (see Chapter 11, footnote to Section 11.1.1), and my talk at the Saltzburg conference in 1968, published as B. de Finetti, 'Initial probabilities: A prerequisite for any valid Induction', *Synthese*, XX, 1 (1969), are devoted to a refutation of this, and to the clarification of various problems connected with it.

Related topics were mentioned at the conference by Vetter, Hintikka, von Kutschera and Frey; in particular, see the paper by I.J. Good, 'Discussion of Bruno de Finetti's paper', which reveals the differences in attitudes existing within the subjectivistic conception.

(and there have been attempts to put forward abstract theories of 'methods of measurement' at this level).

The methods proposed by objectivistic statisticians are at an intermediate level. The probabilistic framework is accepted for that which takes place conditional on certain given hypotheses, but any reference to a probability distribution for the hypotheses themselves is rejected. To relate this to our previous considerations, the 'hypotheses' are the various values of the parameter θ and what is rejected is the prior distribution $\pi_0(\theta)$ (and hence the posterior $\pi_0(\theta|x)$).[13] All that one is permitted to work with is the assumption that the X_h are stochastically independent with the same distribution, $f(x|\theta)$, conditional on each value of θ.[14]

The implication of this for problems of estimation (and similarly for 'tests of hypotheses') is that the function ϕ can only be made to depend on the $f(x|\theta)$, whereas in the unrestricted (i.e. Bayesian) formulation one must also make it depend on $\pi(\theta)$.

12.6.3. One way of avoiding the difficulty is to use the Bayesian approach (either consciously or unconsciously) but omitting $\pi(\theta)$: in other words, by implicitly adopting the (possibly improper) prior $\pi(\theta) = \text{constant}$. In this way, the conclusions obtained are necessarily valid, although it should be noted that indiscriminate use of this prior may result in its adoption in situations where neither the individual using it, nor the majority of other people, find it reasonable. Worst of all, actual contradictions can arise if the approach is used independently in related problems (as a trivial example, taking first θ to have a uniform prior and then, later in the same problem, taking $1/\theta$ to be uniform).

One way of running head on into the difficulty – whilst claiming at the same time to have solved the problem – is to assert that nothing can be said concerning the probability of the statement of interest being true (e.g. that x_0 is 'close to $\hat{x}$', or that it lies between x' and x''). Having decided against overcoming this problem by the use of prior probabilities, it suffices … to pretend that the solution we require is, in fact, a different one, and concerns the probability of the statement of interest being true conditional on the (false!) hypothesis that x_0 is known. In fact, the statements would be similar in appearance only. We could gloss over it by saying, in either version, that '*in any case*, it is almost certain that x_0 and $\hat{x}$, the true and estimated values respectively, are close to one another', as if the phrase 'in any case' had some abstract and absolute meaning, both when it refers to 'whatever the true value might be' and when it refers to 'whatever the estimated value might be'.

The fallacy in confusing the two cases is obvious. In fact, under the usual assumptions, it is almost certain that the mean of the measurements obtained from n observations will turn out to be near the true value (whatever it may be), since we are assuming the error distribution to be the same no matter what the true value is. When we consider the mean resulting from a set of given observed values, however, we are by no means entitled to conclude that the true value is almost certainly near this mean. It may well be that, finding the latter conclusion hard to believe, one considers it as much more

13 There is no objection, however, in problems where objectivists adjudge there to be an 'objective' prior. In such cases, the approach will be the same as a Bayesian would adopt.

14 There seems little point in complicating this brief account by extending it to include the more general cases (e.g. those with $f(x|\theta, y)$, and so on).

plausible (even though *a priori* quite improbable) that, by chance, the observations have turned out to be affected by large errors acting in that particular direction.[15]

12.6.4. There are critics who occasionally attempt to ridicule this argument by pretending to interpret it as meaning that the difference between x and x_0 can be small at the same time as that between x_0 and $\hat{x}$ is large.[16] As we have stated above, we are not drawing a distinction between these two cases (there is none) but between conditioning on the hypotheses 'whatever x_0 may be' and 'whatever $\hat{x}$ may be'. Tracing this back to Bayes's theorem, what goes wrong is that those who do not wish to use it in a legitimate way – on account of certain scruples – have no scruples at all about using it in a manifestly illegitimate way. That is to say, they ignore one of the factors (the prior probability) altogether and treat the other (the likelihood) as though it in fact meant something other than it actually does. This is the same mistake as is made by someone who has scruples about measuring the arms of a balance (having only a tape-measure at his disposal, rather than a high precision instrument) but is willing to assert that the heavier load will always tilt the balance (thereby implicitly assuming, although without admitting it, that the arms are of equal length!).

These same comments apply, essentially unaltered, to the case of hypothesis testing and to other topics (and so we shall not bother to repeat them), because they relate to the essence of the whole 'objectivistic' approach to statistics. One important consequence is the realization that objectivistic forms of significance test do not obey the likelihood principle. These are tests in which, for example, one rejects the null

---

15 If we call X the true value, Y the estimated value and $Z = Y - X$ the error, it is clear that the distribution of Z given $X = x_0$ is not the same thing as the distribution of Z given $Y = y_0$. If $f(x, y)$ represents the joint density for (X, Y), then, in the two cases, the distributions of Z are given by $Kf(x_0, x_0 + z)$ and $Kf(y_0 - z, y_0)$, respectively. These can only coincide if X and Y both have improper uniform densities and Z is independent (i.e. $f(x,y) = Kg(y - x)$ with $K = 0$, in the usual sense). In the case of Section 12.3.2 (X and Z independent and normally distributed), $f(z|x)$ is independent of x by hypothesis (normal distribution $N(0, \sigma/\sqrt{n})$ but $f(z|y)$, as is shown in (16), although still normal and having the same variance, has nonzero prevision:

$$\mathbf{P}\left(Z \mid Y = y_0\right) = m_f - y_0 = \left(m_0 - y_0\right) / \left[1 + n\sigma^{-2} / \sigma_0^{-2}\right],$$

where y_0 was denoted in equation 12.16 by $n\bar{x}$. The term $m_f - y_o$ only vanishes if we take $\sigma_0 = \infty$; i.e. the improper uniform distribution over $\pm\infty$.

Objectivists will probably argue that, as a rule, really large errors do not occur and that if they do one notices the fact and rejects the observation. However, the rejection of a complete, coherent formulation cannot be justified under the pretext that if something does not seem to work one can always get out of trouble by resorting to expedients which themselves cannot be justified (neither in the new, patched-up formulation, nor in the coherent one).

16 This is a rather imprecise objection, open to several interpretations. Only laymen (so far as this topic is concerned, anyway) could take it literally as providing evidence of an oversight. That we are dealing with two different things (see the explanation in the text) is clear, not only to the Bayesian but also to objectivists of the Neyman–Pearson school. The difference is that the latter deliberately choose to base themselves on considerations of the form 'whatever x_0 may be', in order to avoid the Bayesian formulation (assuming arguments based upon the former considerations to be valid, despite the fact that they do not have the same meaning as those in the Bayesian framework, and, indeed going so far as to claim the former as 'modern', and the latter as 'old fashioned'). R.A. Fisher, on the other hand, attempted to create a fusion of the two. It seems to me that he felt the need for the Bayesian form of conclusion (although he expressed it in an illusory manner by means of an undefinable 'fiducial probability'), but wanted to approach the problem from the opposite direction (an approach rather like that of Neyman).

hypothesis $\theta = \theta_0$ because some given function $t(x)$ of the observed data (a statistic) has 'too large' a value (lying outside some given confidence interval; i.e. in the 'tails' of the distribution of t). The point concerning the likelihood principle is clear, because, for the objectivist, the confidence interval is one in which, with $100\beta\%$ probability, $t(x)$ must lie given θ_0 (and not vice versa!). An example of this is given in Lindley, Vol. II, pp. 68–69.

These strictures do not imply, however, that the conclusions cannot, in practice, be satisfactory for most applications. Referring to our example, it will, in fact, be very rare for x_0 not to be close to $\hat{x}$. However, why should we blind ourselves to the possibility of it being otherwise? Why should we stick to the standard conclusion even in cases where we are suspicious? Why should we be forced to ignore facts which, if we do not wish to shut our eyes to them, should lead us to be suspicious?

In any case, a Bayesian analysis will indicate within what limits, and under what conditions, any particular method is approximately valid and what needs to be done (following Lindley's example, perhaps) in order to turn it into an exact and acceptable procedure.

12.6.5. The method of *maximum likelihood* was developed in particular by R.A. Fishe, and, although it was known previously, it was through his work that it came to prominence.

In its crudest form, as a method of point estimation, it consists in taking the estimate of a parameter θ as the value (or vector etc.) $\hat{\theta}$ that gives the (absolute) maximum of the likelihood for θ given by the observations x. One can give this a Bayesian interpretation as the estimate of θ given by the *mode* of the posterior distribution, assuming the prior to be uniform (since the posterior *coincides*[17] with the likelihood in this case, the point maximizing the former also maximizes the latter).

The most useful application of the concept is in providing a *normal* approximation to the posterior distribution (which can then, if one wishes, be used to give an interval estimate).

If we consider the standard case of exchangeability, that is repeated observations with the same density $f(x|\theta)$, the likelihood is the product, as we have seen many times before, and its logarithm is given by

$$L_n(x\,|\,\theta) = L(x\,|\,\theta) = \log p(x\,|\,\theta) = \sum_{h=1}^{n} \log f(x_h\,|\,\theta). \tag{12.31}$$

The logarithm is used simply for convenience and the function L, to use the standard notation, is called the log-likelihood (the subscript n usually being omitted).

As n increases, the influence of the prior distribution π_0 on the posterior π_n becomes smaller and smaller, the fixed factor being overwhelmed by the n factors of the form $f(x_h|\theta)$. This was clear even in Section 12.2.5 (see equation 12.3′) and especially so in the examples we studied (see equation 12.15 of Section 12.3.2 etc.). This means that the more observations that are available, the more their influence is predominant in

17 Recall that it is not really correct to say 'coincides', because the likelihood is a point function, whereas the density depends on the point *and* on the measure (see the final remark of Section 12.3.4).

determining our posterior opinions and the less significant the prior opinion becomes. This is what we would expect.

Because of this (a fact which, incidentally, has been appreciated for quite a while, and was well illustrated by Poincaré), the difficulties we mentioned relating to the evaluation of the prior probabilities turn out to be less serious from a practical point of view. We are not saying that the problem disappears, but that it becomes possible to deal with it in a satisfactory manner by making precise the conditions and the limits within which it is possible to replace a given prior distribution by the uniform, for example, without causing any serious distortion.

In general, and under fairly weak conditions, the likelihood, for large n, is sharply peaked around its maximum, so that the maximum likelihood estimate $\hat{\theta}$ is, as it stands, quite informative (and this is true, in particular, in the case of the normal distribution). A point estimate on its own, however, is never very satisfactory and it is fortunate that the maximum likelihood approach enables us to improve on this by also providing the variance, not of $\pi_n(\theta)$ itself, but of the normal approximation to it in a neighbourhood of $\hat{\theta}$. In fact, we have

$$\sigma_n^{-2} = -\frac{\partial^2}{\partial\theta^2} L_n\left(x \mid \hat{\theta}\right). \quad [18] \tag{12.32}$$

A rough argument will suffice to show why this is so:[19] expanding $L_n(\theta) = L(x|\theta)$ around $\theta \simeq \hat{\theta}$, we obtain

$$L_n\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right) L_n'\left(\hat{\theta}\right) + \frac{1}{2}\left(\theta - \hat{\theta}\right)^2 L_n''\left(\hat{\theta}\right) + \ldots,$$

and the density (again, in a neighbourhood of $\hat{\theta}$) is given by

$$\pi_n\left(\theta\right) = Kss\pi_0\left(\theta\right) \mathrm{e}^{L_n(x|\theta)} \simeq K.\,\mathrm{e}^{+\frac{1}{2}\left(\theta-\hat{\theta}\right)^2 L_n''\left(\hat{\theta}\right)}, \tag{12.33}$$

because: (i) $\pi_0(\theta)$ (in the small neighbourhood of $\hat{\theta}$ in which $L_n(\theta)$ is large) is practically constant (and equal to $\pi_0(\hat{\theta})$); (ii) $L_n(\hat{\theta})$, which is constant, can be subsumed in K; (iii) $L_n'(\hat{\theta}) = 0$, because L_n is maximized at $\hat{\theta}$; and (iv) we can neglect terms beyond those of second order.

The term $L''(\hat{\theta})$ is called the 'information' (but must not be confused with the concept as used in information theory; see Chapter 3, 3.8.5). The result can be extended to the case where θ is a vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$. The distribution is then multivariate normal and a natural generalization of equation 12.32 defines the *information matrix* as the inverse of the variance-covariance matrix:

$$I_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} L\left(x \mid \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s\right). \tag{12.34}$$

---

18 Here, and in equation 12.34, the derivative of $L_n(x|\theta)$ is evaluated at $\theta = \hat{\theta}$.

19 This is basically the same argument as that given in Chapter 11, 11.4.4.

In other words, the I_{ij} give the coefficients of the $(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$ terms in the quadratic form $-Q$ appearing in the density $K.\mathrm{e}^{-\frac{1}{2}Q}$.

12.6.6. Although our account has been very brief, it has dealt with several of the most important topics of mathematical statistics, both from the Bayesian and the objectivistic standpoints. Moreover, we have indicated the main points of departure of the two approaches.

In particular, we note that, in practical terms, the situation is altered by the fact of whether n is large (when we enter the realm of so-called *large-sample* theory) or small (*small-sample* theory). In the first case, practically any method works; whereas, in the second, different methods lead, in general, to very different conclusions. The Bayesian approach is part of a coherent, formal theory, which rules out any conceptual obscurity. On the other hand, in the case of small samples the conclusions are strongly dependent on prior opinions, which may vary greatly from one individual to another. This is a genuine unavoidable fact, but it is not a drawback of the Bayesian approach. It would be a drawback if it were an unnecessary complication but the fact is that if complications do actually exist the drawbacks and errors stem from ignoring them and providing pie-in-the-sky solutions that do not take them into account (as in the objectivistic approach).

Anyway, in concluding this summary I should like to quote the following words of Lindley (Vol. II, Preface, p. xii):[20]

> 'Most of modern statistics (*i.e. that of the objectivistic school*) is perfectly sound in practice; it is done for the wrong reason. Intuition has saved the statistician from error. My contention is that the Bayesian method justifies what he has been doing (*by reinterpreting and correcting it*) and develops new methods that the "orthodox" approach lacks.'

12.7 The Connections with Decision Theory

12.7.1. It is not our intention to discuss this topic at all thoroughly, nor would it be possible to do so within the limits of the present outline treatment. Had we wished to do so, however, we could have set the whole of this chapter within a decision-theoretic framework. What we shall do is to clarify some of the areas in which decision theory offers additional insights into certain of the problems of mathematical statistics and into the comparison between the Bayesian and the more fashionable objectivistic approaches.

We mentioned this topic briefly at the end of Chapter 3 (and here and there in the sequel), where we observed that coherence required us to adopt the criterion of maximizing (expected) utility as the basis of decision making.

Basically, it tells us that we should arrive at a decision by first considering the individual increments of utility attached to the consequences of the various possible decisions, and then weighting these by the respective probabilities. A decision must, therefore, be based on probabilities; that is the posterior probabilities as evaluated on the basis of all

20 The explanations in parentheses, together with the quotation marks for 'orthodox', are not part of the original.

information so far available. This is the main point to note. In order to make decisions, we first require a statistical theory that provides conclusions in the form of posterior probabilities. The Bayesian approach does this; other approaches explicitly refuse to do this.

Indeed, objectivistic approaches to statistics bend over backwards to give nonprobabilistic answers to probabilistic questions, expressing them in YES–NO terms, as in the logic of certainty. More specifically, they talk in terms of 'accepting' or 'rejecting' a given hypothesis on the basis of some given test and, although some hesitate to go this far, occasionally one hears that 'to accept an hypothesis' means 'to agree to behave as if it were certainly true'. This is nonsense. One should not behave 'as if an hypothesis were certain' unless it actually is regarded as certain. If it is not, then we cannot decide how to behave until we have attributed to it some probability p. The appropriate behaviour is then that which, on the basis of p, is calculated to maximize expected utility.

12.7.2. Some authors (notably R.A. Fisher) criticize the application of these ideas to problems of scientific inference, regarding them as essentially economic in nature and incompatible with pure research. We could object that even in the scientific field one cannot escape having to weigh up favourable and unfavourable consequences, but a more decisive reply stems from the fact that these 'economic' arguments reveal the necessity of making sure that opinions cohere. In particular, they show that one must pass from prior to posterior opinions in conformity with Bayes's theorem, and that this is the case no matter whether we are contemplating a bet, a business decision, or simply recording our conclusion for use in a scientific context.

There do not exist two entirely different forms of valid reasoning, one suitable in a commercial context, the other for pure research. No one working in the scientific field considers it beneath him to use the same arithmetic operations, or calculating machines, as are needed for commercial purposes. There is only one theory and it does not matter whether it is used for utilitarian purposes, or for pure research, or simply studied for its own sake.

12.7.3. It is interesting to note that the movement towards a decision-theoretic point of view began within the framework of objectivistic theory. Above all, this was the result of Abraham Wald's introduction of the idea of associating a loss with an incorrect decision, taking, as an example of this, the acceptance of an hypothesis i given that hypothesis j is true (loss $= L_{ij}$, zero if $i = j$). However, this does not entirely remove the unsatisfactory identification of the decision as the 'acceptance of an hypothesis'. The necessary step involves singling out the individual possible 'actions' – choice among which corresponds to a probabilistic assessment – rather than acceptance of the various hypotheses. Some criteria of decision making are taken over from other contexts, without examining closely their suitability for the problem under consideration (for example, the minimax criterion is considered acceptable, even though it corresponds to a different situation, that of competitive uncertainty – i.e. games theory). However, Wald's formulation did result in the explicit reintroduction of prior probabilities and hence Bayesian theory (albeit in a formal sense, without involving the subjectivistic interpretation).

Other movements in this direction have sprung from criticisms of various paradoxes and defects within the objectivistic framework itself. In order to remove these, it became clear that a Bayesian formulation was required. In this context, the contributions of I.J. Good, D.V. Lindley and L.J. Savage deserve explicit mention.

In addition there has been a great deal of research into the economics of uncertainty. Through the work of von Neumann and Morgenstern this gave new life and impetus to the study of utility theory, which had long been neglected (although various scholars – Daniel Bernoulli in the past, and F.P. Ramsey more recently – had shown an interest in it). These various strands of research found their culmination in the work of L.J. Savage, *The Foundations of Statistics* (1954) (so far, that is, as the theme of this book is concerned; the revision of statistical methodology from a Bayesian point of view is more recent, and is still continuing).

12.7.4. Finally, we should note that decision theory has very important things to say about questions relating to the planning of experiments for statistical purposes. In other words, planning experiments in order to improve the information on the basis of which decisions are to be taken.

One aspect of this involves the techniques of such experiments, these being studied in order to optimize the outcome; that is to obtain the most valid and useful information at the least possible cost. It would take us too long even to just mention the most important problems and methods, which have been extensively studied in the literature. It will suffice to simply point out that a vast amount of research has been done and that its enormous contribution to technological progress cannot be properly appreciated unless one examines a number of examples.

So far as we are concerned, it is the more basic aspect of all this which interests us. We are referring to the fact that the reasons which make clear to us the correct form of argument for reaching useful, practical decisions, and that required for reaching conceptually valid conclusions, are the *same* in both cases.

In fact, the seemingly 'new' problem, '*what information is it most useful to obtain before making the decision*?', can be considered as it stands within our previous formulation, underlining yet again its general and comprehensive nature. It suffices to include as possible 'actions' not only those of the original formulation – that is those relating to the 'final decision' – but also the various possible choices of experimental procedure and model-building which lead up to it. The value of any piece of information (in the context of a particular decision problem) can be measured as the increment of expected utility deriving from it (or, in the simplest case, the increment of expected gain). This value is always positive (although it could be zero; if the worst comes to worst, we can always take the decision without taking into account the additional information) but there is usually some cost incurred (for labour, time etc.). The net gain from the information (or, more precisely, from the decision to obtain it) is the difference between its value and its cost. The optimal decision (regarding what information one should seek to obtain) is given by that for which the difference between the value and the cost is maximized. In general, the process of collecting information may be quite complicated (performed sequentially, in a number of stages, with a built-in arrangement for subsequent choices to depend on the information obtained initially, and so on). From a conceptual point of view, our general approach can cope with all this without requiring any modification.

In this context, it is clear from a practical point of view that there is a need for coherence not only for each individual decision but also at an overall level, linking the individual steps together. Such a requirement is perfectly obvious if problems are set out in a detailed fashion within their natural probabilistic setting, but it tends to be overlooked if one gets used to dealing with problems on a fragmentary basis.

Typical of the confusion that can arise is the statement that the 'minimax' procedure (in decision theory) is coherent. In actual fact, it is coherent for each individual application, because it turns out to be *Bayesian* under the choice of a particular prior distribution (and any one is free to choose it if they wish). It corresponds to the choice of the *least favourable* distribution, one that would be used by an opponent who wished to make things as difficult as possible for us. This analogy with games theory – more precisely, with two-person zero-sum games, that is those in which one person's loss is the other's gain – is often emphasized by referring to a statistical decision as a 'game against Nature'. The analogy only goes through, however, if one assumes 'Nature to be malevolent'.

Apart from any reservations one might have about this latter hypothesis,[21] we see at once that it cannot be applied in every case. In fact, if we simultaneously consider a number of decisions all depending on the same event, this approach will certainly lead to contradictions, because the least favourable distribution for one decision will not, in general, be the least favourable for the others. Nature (nor any other opponent for that matter) cannot be so evil-minded as to simultaneously adopt distributions – or '*strategies*' as they are called in games theory – which necessarily put us in a least favourable position for *any individual* decision problem that we might wish to consider.[22] As an obvious analogy, anyone being pursued by a number of hunters coming at him from different directions cannot escape in the opposite direction to all of them.

12.7.5. It might appear that these, our final considerations, have only been made possible by the long and wearisome journey that has gone before. In fact, this is not so. If one sticks to the approach that we have advocated throughout, all this – and let us repeat it once more, so that there is no doubt – is obvious. The time and energy was required for the long excursion that we made into objectivistic territory – a necessary journey, undertaken not as an end in itself, but in order to dispel the notion that an objectivistic formulation could constitute an acceptable, alternative approach. That journey is now over and our work is done. Free at last from paradoxes and contradictions, we emerge from our sea of troubles.

---

21 Some people attempt to justify it as a 'conservative policy' for anyone wishing 'to guard themselves against the risk of the worst happening to them'. The solution to this, if there is one, lies in choosing a very convex utility function, not in deliberately distorting one's opinions; this can only result in a worse decision, and is therefore unacceptable.

22 Anyone wishing to take seriously the hypothesis that Nature is ill-disposed towards him, should adopt a prior distribution that is least favourable *over the whole range of decisions* confronting him, and involving the circumstances under consideration. This would involve applying the minimax criterion to the single, compound problem (taking the entire complex of possible decisions as a single decision), or, alternatively (if the cases are independent), solving each individual decision one at a time, but in terms of what 'Nature's evil-minded strategy' would be over the whole complex of decisions (a very different situation from individual applications of minimax). This is the same as the distinction between minimizing a sum of functions $f(x) = \sum_n f_n(x)$ – i.e. $f(\xi)$ at the value ξ where the sum obtains its minimum – and evaluating $\sum_n f_n(\xi_n)$ the sum at the individual minima (as if x could assume simultaneously – perhaps being evil-minded – the different values $x = \xi_n$).

Appendix

1 Concerning Various Aspects of the Different Approaches

In every field, and in particular in the calculus of probability, there is scope, both hypothetically and in fact, for a number of axiomatic approaches, each of which, to a greater or lesser degree, differs from the others in various respects. It does not suit our purpose to choose just one of these, merely illustrating – even if exhaustively – that particular one; nor are we interested in presenting a somewhat wide and diverse collection from which each person makes his choice with the aid of a pin. The way that seems more appropriate, and that in any case we shall try to follow, consists in sticking to one preferred approach as a reference point but at the same time illustrating both the variants within it that seem admissible, or necessary, and the approaches inspired by divergent views. This provides the framework for the necessary conceptual and formal comparisons.

From a conceptual standpoint our choice has already been made, and explained at some length, in Chapters 3, 4 and subsequently. At that time, we gave what might be called an axiomatic approach, but between then and now there is a difference in attitude that can be expressed (in the summary form of a single sentence) by saying that we must pass from an axiomatic approach to the *theory of probability*, to an axiomatic approach to the *calculus of probability*. This transition must not be taken as implying the existence of any distinction or separation between the two terms, or the desirability of creating such a distinction; we simply wish to draw attention to the different perspective that is obtained by emphasizing on the one hand the essential meaning, on the other the formal aspect.

The difference in perspective is the same as that which occurs when a given theory is viewed by a physicist and a mathematician. One concentrates his attention on the passage from the 'facts' to their mathematical translation; the other on the work involved in the latter. This then resolves itself into the difference between the axiomatization of a theory considered *from the point of view of its meaning*, and the axiomatization of a theory reduced *to its formal and abstract aspect*. In the first case, with reference to the example of a physical theory, the *axioms* encapsulate all those properties of an experimental nature that have been ascertained (or are assumed, perhaps hypothetically, to have been ascertained), and which suffice to give meaningful (i.e. operational) definitions of concepts and quantities, and to establish a mathematical theory to which they are subjected. In the second case, however, we omit the details and merely assume

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

the result as our starting point: the *axioms*, independently of the meaning and validity of the physical interpretation from which one starts, are now nothing more than an expression of the mathematical nature of certain entities and of the form of the relations among them. In this way, the mathematician can work with the axioms without worrying about those features which do not concern him *qua* mathematician. As always, the division of labour carries with it both advantages and disadvantages. A blind man with very acute hearing and a deaf man with very sharp eyesight will be able, in conjunction, to see and hear better than a normal individual, but they might 'understand' less owing to their inability to communicate. We will return to this point later.

The distinction we have just considered applies equally to the case of probability. As an axiomatic approach to the *theory of probability*, we understand the axiomatization made from the point of view of *meaning*. The latter consisted, for us, in the analysis of the conditions of coherence for bets (or something similar) on things we called 'events'; for others, it may consist of assertions either about symmetries, or frequencies, or things also called 'events', but which, perhaps, might be thought of as 'sequences of events', or whatever. In this way, one comes to impart meaning to certain words (quantities etc.) and to establish relationships that must hold among them. As an axiomatic approach to the *calculus of probability*, we mean the axiomatization made from the *formal and abstract* point of view: we have rules with which to operate on symbols without the necessity of knowing which, if any, interpretation these rules and symbols have in the actual context.

Of course, such a contraposition is too crude to serve as anything other than a starting point; on no account must we gloss over the finer points (perhaps hidden to a superficial view, but nonetheless essential). In the choice of the mathematical axiomatization there is plenty of scope for choosing among formulations that are formally equivalent (but whose particular axioms, to those who recall the original meaning, might differ in their intuitive appeal); on the other hand, choices that are made concerning the more 'peripheral' aspects can appear rather arbitrary and made simply for mathematical convenience.

The path we shall follow is motivated by our steadfast refusal to adopt this bad habit. In precise terms: *the axioms of the calculus of probability will be nothing more, and nothing less, than the translation into an abstract form of the conclusions which follow strictly from the practical exigencies brought to light during the preliminary discussions concerning the theory of probability*. It is useful, at this point, to clarify, in a summary and preliminary fashion, why this statement, so obvious in itself, is, instead, at odds with all those formulations, which, by following the same criterion *a little less strictly*, end up, in my opinion, by not following it at all. These clarifications will certainly not be enough to give a sufficient picture of the many factors to be taken into consideration and of their compass. However, they will enable those who bear them in mind to get to grips with the many considerations that will have to be worked out in detail, but without repeating too often, and tediously, these general motives.

We know that what we have to deal with in any case will be the characterization of certain functions **P** defined over the field of entities E, called 'events' (and then **P** is called 'probability'), or over the wider field of entities X called 'random quantities' (and then **P** is called 'prevision'). In order to carry out our task we will try to pose the formal questions concerning events in such a way as to reproduce, as faithfully as possible, the circumstances that can practically arise for events (together with variants – some

important, some less so – to meet particular exigencies): similarly for random quantities. In order to define the functions acceptable as **P**, we will utilize only the conditions of coherence expressed in an abstract form.

In what way does this differ from the formulations more usually adopted at the present time? In the first place, the structure which is generally preferred is a closed, monolithic one. Rather than defining events in a general way, and then the functions **P** as extendible (in principle) to all events (either already conceived of, or conceivable in the future), one constructs on each occasion a definite, well-delimited (although possibly enormous) field of events with a particular function **P** attached to it once and for all. In terms of the standard image (in which events are thought of as sets in an abstract space), this means that the complete set-up (or 'probability space') is a *measure space* (i.e. a space with *one* particular, fixed measure). In contrast to this, the separate consideration of first the *space* (*without the measure*, or any other kind of structure) and then all the possible *measures*, not only, and most importantly, meets the needs of the subjective conception by providing $\mathbf{P}_i$, which are possibly different for each individual i ('tot capita, tot sententiae'), but also satisfies other more 'neutral' requirements (probabilities conditional on different hypotheses, or different states of information, or 'mixtures', and so on).

Moreover (independently of the previous objection, concerning **P**), this space–measure coupling gives rise to an unnatural, forced relationship between the two notions of event and probability, because it does not take account of the problems raised by the fixing of a particular function **P**. The current practice of reducing the calculus of probability to modern measure theory (countably or σ-additive, as in the Lebesgue theory) – apart from changes in terminology (set–event; measure–probability; function–random quantity; integral–expectation) – has resulted in the following:

- probability is obliged to be not merely additive (as is necessary) but, in fact, σ-additive (without any good reason);
- events are restricted to be merely a subclass (technically, a σ-ring with some further conditions) of the class of all subsets of the space (in order to make σ-additivity possible, but without any real reason that could justify saying to one set 'you are an event', and to another 'you are not');
- people are led to extend the set of events in a fictitious manner (i.e. not corresponding to any meaningful interpretation) in order to preserve the appearance of σ-additivity even when it does not hold (in the meaningful field), rather than abandoning it.

Among other things, in the case of limiting processes and definitions of stochastic limits, this leads to the adoption of formulations that are unacceptable as they stand unless σ-additivity is imposed (at the cost of a great deal of inconvenience) as a necessary assumption at all times.

We should, of course, discuss these objections and reservations rather more fully, and go on to justify them; all the more so in that they will seem strange to those who are accustomed to the standard formulation. In fact, in the standard approach the points which do not seem to stand up to a critical examination are introduced either with the tacit suggestion that they are obvious, or they are couched in suitably seductive terms to overcome any initial reluctance to accept them.

There are other negative features of the space–measure coupling which are not related to the assumption of σ-additivity. An example is provided by the fact that *zero probability* is regarded as a property of the event in question (among other things, this sometimes

leads to two events, $A \neq B$, being defined as 'equivalent' if their symmetric difference, $A\tilde{B} + B\tilde{A}$, or $A + B - 2AB$, has zero probability). Even more dangerous is the fact that *stochastic independence* – $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$ – is considered as being a property of the events; and so on. One should beware of laying insufficient stress on the fact that it is a property of the function **P** (in relation to the events A and B) and not of the events as such (but this important distinction ceases to have any meaning if **P** is considered as given!).

In Chapter 2, we gave an account of what can be said about events from a logical standpoint (within the logic of certainty); in other words, concerning the events in themselves. We postponed until Chapter 3 anything which depended on the introduction of the function **P** defined on the events (without adding or altering anything concerning the notion of event, or the meanings of individual events). This separation was made in order to avoid any confusion early on; confusion which could have led to misunderstandings later.

Here we shall adopt the same policy, although, of course, in a deeper, more systematic and precise way. Certain distinctions that appear meaningful in other formulations no longer appear so in ours. Consider, for example, the distinctions between whether or not events are *atomic* (i.e. contain no events other than themselves and the empty one; in terms of sets of points, this reduces to those sets formed from the singletons), or between those events *belonging to either finite or infinite sets* of events, and so on. In the case of a random quantity X, having as possible values, for example, all real numbers between 0 and 1 (like X = 'percentage of time during which a given telephone link will be busy tomorrow between 9 a.m. and 5 p.m.'), let us consider the event $E = (X = 0{\cdot}4166666\ldots)$. It consists of obtaining exactly a given preassigned value and could be regarded both as 'belonging to an infinite set' (i.e. of events $E_x = (X = x)$, $0 \leqslant x \leqslant 1$) and as an 'atomic event' (because a precise value, like $x_0 = 41\frac{2}{3}\%$, does not admit further refinement). It also belongs, however, to the field consisting of just the two events $E = (X = x_0)$ and $\tilde{E} = (X \neq x_0)$ (together with the events 0 and 1), and can be decomposed into $E = EA + E\tilde{A}$, by means of any event A not involving X (for example: A = 'it will rain tomorrow'; A = 'the party at present in government will not remain in power after the next election'; A = 'the azaleas in the window of the florist across the street will be sold today'). This can be extended to infinite subcases by considering other random quantities $Y, Z, \ldots$ (and, therefore, by considering as 'provisional atoms' the points (x, y), or (x, y, z), or $(x, y, z, \ldots)$ of $S_2, S_3, \ldots, S_n, \ldots$). It follows that any considerations put forward on the basis of these nonexistent distinctions must be without foundation (an example of this is the assertion that an event E that is not impossible can only have zero probability, $\mathbf{P}(E) = 0$, if it 'belongs to an infinite set of events').

On the other hand, there exist real problems that arise in various connections with the notion of the 'verifiability' of an event; a notion which is often vague and elusive. Strictly speaking, the phrase itself is an unfortunate one because verifiability is the essential characteristic of the definition of an event (to speak of an 'unverifiable event' is like saying 'bald with long hair'). It is necessary, however, to recognize that there are various degrees and shades of meaning attached to the notion of verifiability. Some are more or less flexible: verifiable with a greater or lesser degree of *precision*; or within a shorter or longer period of *time*; or with a higher or lower level of *expenditure*; or with a greater or lesser *number* of partial verifications; and so on. Others are more precise: for example, we could consider 'absolute' degrees of precision, or 'infinite' time periods, and so on. The most precise and important, however, is that which arises in theoretical physics in connection with *observability* and *complementarity*. It seems strange that a

question of such overwhelming interest, both conceptually and practically (and concerning the most unexpected and deep forms of application of probability theory to the natural sciences), should be considered, by and large, only by physicists and philosophers, whereas it is virtually ignored in treatments of the calculus of probability. We agree that it is a new element, whose introduction upsets the existing framework, making it something of a hybrid. We see no reason, however, to prefer tinkering about with bogus innovations rather than enriching the existing structure by incorporating stimulating refinements (disruptive though they may be).

It is our intention, therefore, to attempt to provide in this appendix an integrated view of questions of this kind that arise in connection with events. We should perhaps make it clear that our 'attempt' will be mainly concerned with the case of theoretical physics, and will consist of little more than a comparison of the positions adopted by various other authors, plus an indication of which position seems to us to be less open to criticism (as well as being better suited to deal with further problems concerning the verifiability of events).

There are other questions (already mentioned many times in passing) which concern the notion of 'possibility', and further aspects are revealed in cases where, either through haste or oversight (or because of one's own limitations, or because of the impossible nature of the task, or whatever), one has not drawn out all the logical implications contained[1] in the information in one's possession. The result of this is that the set of events considered 'certain' is not *closed* with respect to the logic of certainty.

Finally, we shall turn from the preliminary questions concerning events (and hence the logic of certainty) to the introduction of probability. It is the latter that is for us the real subject matter, the principle character as it were, and the rest is simply the setting of the scene.

We must now pass from the considerations that led us (in Chapter 3) to our basic formulation, to consider the *axioms* which constitute their translation into abstract form. The surest way of avoiding any kind of modification taking place during this translation is to directly express things in abstract form without any alterations. It suffices to preserve additivity and non-negativity. So far as the essential considerations are concerned, this rules out the attributing of a positive price (positive prevision) to a transaction (or bet) that will certainly lead to a negative outcome. From the abstract point of view, this obliges **P** to be such that we can never have

$$c_1\mathbf{P}(X_1)+c_2\mathbf{P}(X_2)+\ldots+c_n\mathbf{P}(X_n)>0$$

if

$$X=c_1X_1+c_2X_2+\ldots+c_nX_n \text{ is certainly} <0.$$

These inequalities (imposed for every finite, linear combination) define (as the intersection of half-spaces) the convex set P of admissible functions **P** (and all that remains to be sorted out are a few details, like the possibility of substituting $\geqslant$ for $>$ in the inequalities, and so on).

---

1 As an example of this, consider the matching problem (with n objects). It could happen that someone does not realize that the case of $n-1$ matchings is impossible (see Chapter 3, 3.8.4), either because he is not capable of arriving at this conclusion on the basis of the information available to him, or because it never occurred to him to doubt that all the values 0 to n were possible. It could also be that he had once known the result, but had subsequently forgotten it; or that he had not really forgotten it, but simply overlooked it at the time in question.

On the other hand, we should point out that in expressing these conditions we have made use of, or at least made reference to, random quantities rather than events. In actual fact, writing $E_1 \ldots E_n$ in place of $X_1 \ldots X_n$ would have given practically the same condition[2] by introducing the X_h (which form a linear space) in an indirect fashion as linear combinations of the events (which do not form a linear space). To start directly with the linear space of the X_h (without giving any particular emphasis to events, which are, in any case, part of that space) not only enables one to deal with the whole set-up in one go but also permits one to emphasize the adherence to the essential meaning.

Proceeding in this way, the axioms directly characterize **P** over its entire field of application: that is both over the field of events – where it can be given the name of *probability* – and over the field of random quantities – where it is called, more generally, *prevision* (or price, if we are dealing with practical situations).

This is a great advantage, not only from a formal point of view but also because of the elegant simplification it provides. One avoids not merely the tiresome complication of having to consider two separate cases but also a whole series of difficulties that stem from the fact that such complications are misleading as well as annoying. In the first place, one encounters a tiresome complication if one wishes to formulate the axioms in such a way as to deal only with events, excluding random quantities. A further complication then arises when one attempts to put right this exclusion and define prevision, taking into account that it has already been defined in the particular case of events, where it is called probability.

The obvious way, and the only possible way, of dealing with the exclusion would simply be to remove it – even though not straightforwardly – by means of some device that puts us back on the straight and narrow. It seems, however, that the first, unhappy step obliges us to continue with it in making the second step. In wishing to consider as a *definition* of prevision some relation connecting it with probability, one is led into an extremely unnatural position. In other words, one makes it appear as though the elementary notion of prevision presupposes a knowledge of something much more complicated and delicate; that is, the probability *distribution* itself. Because it is unnatural, the situation is also dangerous, in the sense that it leads one to think that the definition to be made in this way, *ex novo*, allows a certain element of arbitrariness. In other words, that it requires, or permits, a choice of conventions, which are inspired by considerations of convenience.

In mathematical terms, expressed abstractly, all that we have said reduces to expressing a preference for, and then adopting, the first of the two paths open to us (which we indicate here by quoting the opening sentences of a more detailed description that can be found in Bodiou,[3] p. 5):

i) emphasis on linear functional (Riesz, Bourbaki, L. Schwartz);
ii) emphasis on measure (Borel, Lebesgue, Carathéodory, Fréchet, Kolmogorov).

---

2 If one limits oneself to X_h having a finite number of possible values; from here one could proceed to the general case (with bounded X_h) by means of approximations from above and below.

3 Georges Bodiou, *Théorie dialectique des probabilitiés englobant leur calcul classique et quantique,* Gauthiers-Villars, Paris (1964).

The main thing, however, is not the conclusion we reach – that is the choice itself – but rather the *reasons* lying behind this choice. It is not a question of saying which mathematical formulation has the greatest merit from a mathematical point of view, but rather of saying which provides a means of interpreting most directly those things which are most directly significant, most directly important, and, above all, most directly *observable* (in a conceptual sense).

Our attitude towards the difference between the two approaches to the definition of $\mathbf{P}(X)$ (the prevision of X, usually denoted by $\mathbf{E}(X)$ = the mathematical expectation of X) can be clarified by means of an analogy (which is, in fact, exact, apart from the change in terminology). Given a solid body C, one can define its 'barycentre', $B(C)$, say, and also give an operational method of determining it, without formulae; but it will not normally be possible (nor will it be important) to discover the mass distribution of C. In particular, the notion of 'density' at a point is simply a convention, defined by a limit process which, given the structure of matter (molecules, atoms, particles), cannot, strictly speaking, make any sense. However, it can be said that if we assume the density ρ to be known, as a function of the point P, we are then able to say that the mass of the body, $m(C)$, and its barycentre, $B(C)$, must be given by:

$$m(C) = \int_C \rho(P)\mathrm{d}S, \qquad B(C) = \frac{1}{m(C)} \int_C P.\rho(P)\mathrm{d}S.$$

To summarize: the difference we referred to consists of choosing between those definitions that are direct and intuitive, and those expressed in formulae as in the example above. (Note that in the latter case we require a passage to the limit in order to define density and then, to go back to the body itself, we have to do away with the density by integrating it. If there is any arbitrariness in the definition of the integral to be used, there is always the risk that some error is introduced.)

I find this undesirable habit of making simple things complicated to be very widespread at the present time (it is as if people go looking for trouble – and often they find it). I mention this not because I see it as my business to concern myself with it outside the confines of my own subject, but merely to point out that my noticing it and attempting to remedy it in the field of probability theory does not mean that I only see it as having taken root there. It happens more or less everywhere.

2 Events (true, false, and …)

By definition, an event must either be *true* or *false* (see Chapter 2, 2.3.4). It can be *uncertain* (for us, for the time being) only if, and insofar as, we do not possess the information required for establishing its truth or falsity. The same holds for any random entity; in particular, for random quantities. A 'random' quantity X is a quantity which has a well-determined value x; it could be, however, that we are not aware of what this value is (and it is because of this absence of information that it is, for us, for the time being, uncertain and, hence, random). We can, in fact, limit our discussion to the case of events, because any information concerning X is simply information concerning some event of the form $X \in I$ (where I is any set).

But what does it mean to say that an event is either true or false? Two extreme interpretations would consist in making reference to an 'objective truth' or to 'immediate verifiability'. The latter is unobjectionable but is extremely restrictive: it only holds in situations like that of a quiz where the answers can be found by turning to the next page. Even in this case, however, there are a number of implicit assumptions! We have to exclude the possibility of confusion or bewilderment such as would arise, for example, if every time one turned to the answers one found them different from when one last looked; or found them to be different according to whether one read them with the left eye or the right eye; and so on. Everyone will no doubt agree that these kinds of assumptions are ridiculous but it should be noted that there is no logical reason for regarding them as such. One does so because they conflict with certain 'regularities' that 'objective reality' has accustomed us to. (Dually, from a solipsistic point of view, they conflict with certain 'regularities' that have guided us in our construction of our idea of 'objective reality' – in the image of what appears to us in our maybe-real-world-maybe-dream-world.)

Should we let ourselves be guided by the objective interpretation, the first of the two extremes we mentioned above? Up to a certain point this is inevitable (otherwise we would be forever in the grip of a 'ridiculous' scepticism, as in the examples above). It is necessary, therefore, to be constantly on the alert, with a critical attitude, remembering that many statements that appeared to a 'naïve' objectivism to be undoubtedly meaningful had subsequently to be modified and revised in terms of 'operational' definitions in order for it to be possible to give them a meaning.[4]

But when is 'objectivism' not 'naïve'? Unfortunately, the answer is far from reassuring: 'it is so up until the point when the unexpected occurrence of the contradictions or drawbacks to which it gives rise actually take place'.[5] When this happens, one has to seek a remedy, and this consists in moving as far as we can in the opposite direction. In other words, we cease to think of the 'objective' fact of something being either true or false, but rather of the fact of whether or not we can obtain the information that for us determines whether it is true or not (or, at least, whether there is a possibility' – in some sense or other – of obtaining this information).

This would lead us to regard some events as worthy of the name (since it actually makes sense to ask whether they are true or false) and others as requiring elimination (in that they are bogus – events in appearance only, non-events). If the possibility of a clear-cut separation of the two kinds of events existed (or, at any rate, were assumed to exist – possibly with some appropriate, simplifying hypothesis), there would be no problem. Everything could then remain as before (including the definition of event),

---

4 At this point, in order to avoid confusion and misunderstandings, we should clarify the relationship between subjectivism in the field of probability and subjectivism in relation to knowledge in general. It is sometimes said that 'yes, of course probability is subjective … because *everything* is subjective'. Put this way, however, the statement is not in accordance with the subjectivistic conception of probability and is, in fact, at odds with it. The fundamental point of the subjectivistic conception is that the notion of probability does not refer to something which is a property of the 'outside world' (and it does not matter whether the latter is regarded as an 'objective reality' or as a 'mental construct'). A solipsist, who considered all of so-called 'reality' to be 'subjective', in order to be self-consistent, and to correctly interpret the subjectivistic concept of probability, would perhaps be right in saying, instead, that probability is *objective*. It is objective in the sense that it expresses an autonomous judgment and not something which is bound by 'external' circumstances to be interpreted in the sense of 'as if' (Veihinger's 'als ob').

5 It is almost the same as saying that every individual must be regarded as immortal until he eventually happens to die.

with an additional warning that one should make sure that one really is dealing with events (i.e. with events that make sense – those that are verifiable).

Instead, it seems to be necessary to retain more flexibility. More specifically (although this suggestion might appear to be an unhappy expedient), it will be convenient to use the term 'event' quite freely, without any *a priori* selection and exclusion. In this way, the selectivity can be brought in later, case by case, taking into account the different requirements (sometimes clear-cut, more often vague) that arise in connection with 'verifiability', and interpreting them in the light of appropriate (albeit to some extent arbitrary) schematizations.

It might be claimed that by adopting this approach we are begging the question, since the exclusion of that which *must* be excluded (because it is meaningless) becomes mixed up with the exclusion of that which *can* be excluded should we happen not to be interested in it. It is a fact, however, that our analysis (whether completely satisfying or not) does reveal the case of absolute unverifiability to be a limit-case of something more gradual (and, in a certain sense, 'economic'), involving different degrees of difficulty (of various kinds) in verifying whether an event is true or false. Nothing precludes one from evaluating this degree of difficulty in the light of the meaning and importance that such a verification would produce in practical terms.

One great advantage of proceeding in this way (and one which seems to me indispensable) is that our initial scheme remains the same: it includes all those things that we would have called events prior to embarking on these critical considerations, and permits us to carry out all the usual operations on them. When it comes to introducing a restriction (in a way which corresponds – within the given framework – to certain well-defined reasons), it will be sufficient to specify the subclass of events that one wishes to take into consideration (or to regard as making sense, or as verifiable, or whatever), and which other events one wishes to discard (regarding them as bogus, non-events, or as events whose meaning is unclear, or of little interest, or whatever). It may happen that in some circumstances certain given operations applied to verifiable events lead to verifiable events but that in other circumstances they do not. There are various possibilities of this kind and it is simply a question of noting what actually happens, rather than a theoretical question to be posed in abstract form as being an inherent feature of the concept of verifiability.

We have spoken thus far as if it were merely a question of distinguishing between genuine (i.e. verifiable) events, for which there are just two values, True and False, and bogus events, which are either not events at all, or are 'meaningless', or are 'intrinsically indeterminate'. There are cases, however, in which one discusses events for which there are three possibilities: True, False and Indeterminate (or Meaningless). This situation occurs above all in quantum mechanics in connection with the problem of complementarity and, hence, of indeterminism. Having three possibilities could give rise to a three-valued logic (as, for example, in Reichenbach, 1942).

In considering (in Chapter 4) *conditional events* of the form $E|H$, we were, in fact, dealing with logical entities that could take on three values: True ($1 = 1|1$; i.e. both H and E true); False ($0 = 0|1$; i.e. H true and E false); and Void ($\varnothing = 1|0 = 0|0$; i.e. H false and hence the truth or falsity of E irrelevant). This is precisely the way in which the 'three truth values' of the above-mentioned three-valued logic are formed (with H = 'an *observation* made in order to verify whether E occurs or not'; and only after this does it make any sense to ask whether E is true or false). In actual fact, this reduction of the

problem to the simple and familiar set-up of conditional events does seem to provide an adequate solution; moreover, it is especially satisfying in that it avoids any formulation which might appear to contain the germ of metaphysical infection.

There is no problem if only one considers the meaningful components making up the conditional event $E|H$ to be not H and E, but H and EH (see Chapter 4, 4.4.1). So far as the event E in $E|H$ is concerned, it is immaterial whether we take it to be E, or EH (the minimum possible), or $EH + \tilde{H}$ (the maximum possible), or any intermediate event $E = EH + A$, with $A \subset \tilde{H}$. To ask whether E is meaningful (and if so whether it is true or false), when H is assumed false does not make sense, when considered in relation to $E|H$. In this context, one would be considering the question of whether or not the possible residual part of the sentence made sense, or was true or false. If it were meaningful at all, this would represent the irrelevant A, which is outside the field of interest. One could, however, investigate whether, for *other reasons*, the E in the formulation adopted – that is its residual part A which is irrelevant for $E|H$ – should be considered as meaningful and having interest outside of the hypothesis H. This is a separate problem, which concerns the event A as such, and does not have anything to do with the conditional event $E|H$, into which A enters only by the back door (like b in $5a - 0b + 2c$), or with its three logical values (for which, whatever A might be, there corresponds to $\tilde{H}$ – and hence to A and $\tilde{H} - A$ – always and only the same value; $\varnothing$ = Empty).

Let us now turn to a consideration of the mathematical representation of the field of events. We shall mention several variants, their appropriateness depending on the situation and on what is required. When it seems useful to do so, we shall also mention other possibilities that do not fit into our general framework of ideas. We have already provided a great deal of discussion in the text (Chapters 1–12) concerning the reasons for conflicting views on this subject; a few brief comments about certain peripheral topics should therefore suffice here. We shall try to give accurate accounts of those formulations that are not acceptable as such in terms of the approach adopted in the present work, and to bring out their worthwhile features, indicating how they might be applied in particular situations or as special cases.

3 Events in an Unrestricted Field

The basic set-up with which we shall begin is that already described in a summary fashion in Chapter 2, and which we have adhered to ever since, despite the occasional reservation. It serves our purpose in two ways: firstly, it is useful as it stands, in that it provides for a suitable representation and interpretation of the case that we shall regard as the most general, and, apparently, the simplest; secondly, with appropriate modifications, it provides a means of obtaining schemes for representing the other cases of interest (and some aspects of these might, in fact, appear simpler than the first case).

The simplicity of this first case lies precisely in the fact that no restrictions of any kind are imposed when it comes to forming the events: the latter could always be thought of as *arbitrary subsets* of a set of 'elementary possible cases' of a partition admitting indefinite refinements. In formal terms, we could express this more precisely as follows: 'at any given moment', the field $\mathscr{E}$ of events under consideration corresponds to the entire

collection of subsets (or subdivisions) of the set (or *partition*) $\mathcal{Q}$ of the 'elementary possible cases', Q, which, 'at that moment', one wishes to single out. The partition $\mathcal{Q}$ must be considered as having no structure whatsoever and, moreover, it must be considered as 'provisional', 'not once-and-for-all' (and this is what the references to a 'given moment' etc. are intended to convey). This implies that we can only consider as meaningful those notions and properties which are, in a certain sense, invariant with respect to 'refinements' of $\mathcal{Q}$. In more precise forms the latter are represented by subsets $\mathcal{Q}$, for which each set of $\mathcal{Q}$ consisting of a single point Q is replaced by a set containing many 'points' Q' (in general, there could be an infinite number). To summarize: we can provisionally identify the events E of $\mathcal{E}$ with the subsets $\mathfrak{P}(\mathcal{Q})$ of $\mathcal{Q}$, and also with the corresponding subsets $\mathfrak{P}(Q')$ of Q'; – that is without taking too seriously the temporary interpretation of the $Q \in \mathcal{Q}$ as 'points'.

We shall give an example straightaway, in order to clarify this. Let E be the event $X^2 + Y^2 \leqslant a^2$ (where X and Y are random quantities). If $\mathcal{Q}$ is the (x, y)-plane, the event E corresponds to the disc of 'points' (x, y) with $x^2 + y^2 \leqslant a^2$; whereas if $\mathcal{Q}'$ is the three-dimensional space of points (x, y, z) (or four-dimensional, (x, y, z, t) etc.) the event corresponds to the cylinder of points (x, y, z) (or x, y, z, t) etc.) such that $x^2 + y^2 \leqslant a^2$. By considering not only the random quantities X and Y, but also others like Z and T, etc., we change the field $\mathcal{Q}$ and the notion of point (to each point (x_0, y_0) there corresponds the infinity of points on the line (x_0, y_0, z), or on the plane (x_0, y_0, z, t), etc.). The set to which E corresponds – or, conventionally, with which it is considered identified – changes, but this is an irrelevant contingency, arising from the form of representation; what does not change is the meaning of the proposition itself, which is completely contained in the inequality $X^2 + Y^2 \leqslant a^2$.

All this could have been expressed in a better way had we eliminated completely the notion of point, but it appears to be more instructive to put it forward and then to present the arguments against it. In this way, we underline the contrast between the more usual formulations on the one hand, and the refusal to accept 'closure' – as is otherwise inevitable – on the other.[6] On the other hand, it turns out to be useful to accept the 'points' as indicating the limit of subdivision beyond which it is not necessary to proceed (at a given 'moment', i.e. with respect to the problems under consideration). There is just one condition: that we always bear in mind that this is only useful

---

6 To approach the formulation of a theory by starting off with a preassigned, rigid and 'closed' scheme seems to me a tiresome and cumbersome procedure, wherever it is followed. (It is true that it serves to guarantee one against antinomies and suchlike, but this is not a good reason for always having recourse to it; in the same way as it is not necessary to shut oneself inside a tank in order to journey through a peaceful and friendly country.)

In connection with the use of 'points', and their abandonment in geometrical representations, we refer the reader back to our remarks in Chapter 2, 2.4.3 (especially to the quotations of von Neumann and Ulam).

Further discussion, closely relevant to this point, and (insofar as the present topic is concerned) upholding precisely the same position, can be found in Bodiou (1964, p. 3). Abstracting from the space of points, and describing the events directly as the elements of a Boolean algebra, or a Boolean lattice, he observes that 'apart from the formal simplification thus obtained, the new axiomatization is more directly interpretable in terms of the logic of the attributes…. If one assumes that each element of the Boolean algebra is a union of "atoms", one proves equivalence to the Kolmogorov axioms; *the emphasis, however, is more directly on the essential feature – the global lattice, and not the set of its atoms'.* (We should add – although it is not necessary at this point – that Bodiou does, however, retain the usual conditions, which admit σ-additivity.)

insofar as it helps to 'fix ideas' at the time in question. If one were to attribute to it some absolute meaning, it would lead to a confusion of the ideas, and to a tangle of misunderstandings.[7]

The field $\mathscr{Q}$ (or, more generally, a field $\mathscr{S}$ of which $\mathscr{Q}$ is just a part) is often obtained by starting from some given set – which we shall call a basis $\mathscr{B}$ – of events E_h, or, more generally, of random quantities X_h (this was the case in the previous example, where we started with X and Y, and then considered adding in Z and T). To think of the field Q (and therefore of the field $\mathscr{E} = \mathfrak{P}(\mathscr{Q})$) as having been generated from a basis B is completely irrelevant; it is convenient, however, to refer to this case in order to take up the theoretical discussion again, and to develop it in a more expressive manner.

In Chapter 2, and also in the example above, we have already seen how, given n random quantities X_h, the whole picture could be summed up by considering a single 'random point' Q in the Cartesian space $\mathscr{S} = S_n$ (with coordinate system x_h), where Q is the point defined by $x_h = X_h (h = 1, 2, \ldots, n)$. Not all the points of $\mathscr{S}$ are, in general, *possible*, but only those of a subset Q (obtained by eliminating the cases $\mathscr{S} - \mathscr{Q}$ which, on the basis of the data of the problem, turn out to be impossible). In particular, if the X_h are events, $X_h = E_h$, $\mathscr{S}$ reduces to the set of the 2^n vertices of the hypercube (with coordinates 0 or 1), since we can only have $x_h = 0$ or $x_h = 1$. In this case, $\mathscr{Q}$ is the subset of possible vertices; in other words, the constituents (Chapter 2, 2.7.1). In the general case, nothing really changes, except that the E_h, or the X_h, may be infinite in number; the indices h will then run through some infinite set H (not necessarily countable), and even if we write the more familiar $h = 1, 2, 3, \ldots$, or simply say 'all the E_h (or X_h)', we shall mean $h \in H$.

In this way, the preceding (Cartesian) representation will hold without any alteration, except that the number of dimensions (of axes, of coordinates) is infinite,[8] and $\mathscr{S}$ will be S_H (the cartesian space with an infinite number of coordinates, x_h, $h \in H$). In the case of events (i.e. if all the X_h reduce to events E_h) the vertices of the hypercube (in infinite dimensions) are characterized by indicating for which h we have $E_h = x_h = 1$ (for the others, $E_h = x_h = 0$). In other words, they correspond to subsets of $\mathscr{B}$ ($\mathscr{S} = \mathfrak{P}(\mathscr{B})$) or, equivalently, to the functions $f(\cdot)$, elements of $\mathscr{S} = 2^{\mathscr{B}}$, which to some subsets of $\mathscr{B}$ assign the value 1, and to the others 0. One easily recognizes the identical form of procedure to that which led to constituents in the case of finite n; it is a question of stating that out of the events of the basis $\mathscr{B}$, a certain subset are true, and the others false. Of course, some of the products will, in general, be impossible; that is demonstrably false on the basis of the data. We shall need to remove these from $\mathscr{S}$ in order to obtain $\mathscr{Q}$. If we wish, we can always reduce the

---

7 The most serious such misunderstanding likely to arise is the idea that a conditional event $E|H$ has some special significance when H is 'atomic': in other words, when H corresponds to a 'point' in some given representation (although this would obviously be incomplete, since both EH and $\tilde{E}\mathscr{H}$ must make sense, and $H = EH + \tilde{E}\mathscr{H}$). In this way, one would be led to think that $\mathbf{P}(E|H)$ has an absolute meaning, unchanged even if some further information can be added to that expressed by H (here, $\mathbf{P}(E|H)$ is the probability of E, 'knowing all the circumstances that can influence E – and, one might add, determined up to the present moment – as expressed by H').

Considering the early stage we are at in our present attempt at a systemization, the above brief comments may seem premature. However, it is perhaps useful to have some idea of the arguments we shall have to consider, even though we shall only come across them later, while developing this treatment.

8 Note (although this is not really important here) that the concept of the number of dimensions – when this is infinite – could be understood in a different way as the number of nonzero linearly independent elements (and this is actually intrinsically more meaningful): this notion no longer coincides – as in the case of finite n – with the 'number of coordinates'.

case of random quantities to that of events: it suffices to substitute for each X_h the events $E_{h,x} = (X_h = x)$, for all the values x possible for X_h. Calling $\mathscr{B}'$ the modified basis which arises from this substitution for all the X_h, we can always write $S = \mathfrak{P}(\mathscr{B}')$, or $\mathscr{S} = 2^{\mathscr{B}'}$.

If, having constructed $\mathscr{S}$ in this way (using one variant or another), we preserve, even if implicitly, through the x_h, the record of how $\mathscr{S}$ was generated from the basis $\mathscr{B}$, a linear space structure (or that of a subspace) remains as a trace of this in $\mathscr{S}$ (and hence in $\mathscr{Q}$). On the other hand, we might actually be dealing with a problem of geometrical probability (even of geometry, in the sense of ordinary, physical space) and hence we inevitably have the geometric structure (one could think, for example of $\mathscr{S} = \mathscr{Q}$ = surface of the earth, Q = point at which a lost – or stolen – object is located). It does not matter. In saying that Q has to be considered as having no structure, one is not saying that a structure might not be seen to exist if we looked at it *from a different standpoint* (and one does not wish to rule out the possibility of taking this into account if it should appear at a later stage to suit our purpose to do so). We simply mean that *for the time being and for our present purpose* we must ignore it.

If we do not choose to ignore the way in which $\mathscr{S}$ has been derived from the basis $\mathscr{B}$, the possibility arises that we could single out certain events as being somewhat *special*: for example, belonging to the basis, or logically expressible in terms of a finite or countable number of basis elements. Similar distinctions could be drawn among random quantities: those belonging to the basis, or functions of basis elements (the functions being linear, or continuous, or whatever, and involving any particular number of basis elements), and so on. This is why on the real line, starting from the intervals (as basis), one is able to maintain distinctions between sets that are sums of a finite, or countable, number of intervals, or obtainable from intervals by at most a countable number of logical operations, and those which are not. We mention this familiar example merely in order to point out that the introduction of ideas of this kind, and the consideration of such distinctions, is not admissible, and, in fact, must be explicitly excluded, since we wish to regard $\mathscr{Q}$ as having no structure whatsoever (at least for the time being).

There remains just one distinction – a nuisance as far as we are concerned – which, at the present time, would continue to make sense, even if we consider $\mathscr{Q}$ as having no structure. It is that based on the *number* of elements ('points') of the given E and of its complement $\tilde{E}$. More precisely, if the cardinality[9] of the whole set $\mathscr{Q} = E \cup \tilde{E}$ is M (infinite), then either E and $\tilde{E}$ both have the same cardinal, M, or one of them has cardinal M, and the cardinal of the other is smaller (either 0, 1, 2,..., n,..., or $\mathfrak{n}$ – countable infinity – if $M > \mathfrak{n}$, or some other cardinal N, where $\mathfrak{n} < N < M$). The introduction of the convention of 'never' regarding the subdivision as a 'final' one is another reason for ignoring the structural distinctions. In speaking of the *basis*, we can express this by saying that we must always be aware that other events (or random quantities) can be added at will. In this way, the distinction based on the *number* of 'points' also becomes

9 We denote by $\mathfrak{n}$ and $\mathfrak{c}$ (Gothic n and c) the cardinals of the *integers* (the smallest infinite cardinal) and of the *reals* (the continuum), respectively. If A and B are (disjoint) sets with cardinals M and N, respectively, then the cardinals of $A \cup B$ (union), $A \times B$ (Cartesian product of ordered pairs (a, b)) and A^B (the set of functions from B to A), are given by $M + N$, MN and M^N, respectively. If M and N are infinite, and $M > N$, we have $M + N = MN = M^N = M$ (and also $M + M = MM = M$): however, $2^M > M$ (and, *a fortiori*, $N^M > M$), and, in particular, $2^{\mathfrak{n}} = \mathfrak{c} > \mathfrak{n}$ (there are as many subsets of the integers as there are points on the real line).

We shall be making use of these properties in what follows, hence the reason for our recalling them here.

meaningless (and this applies, in particular, to the distinction between 'atomic' events – corresponding to single points – and others). There remains the one structural distinction that we must, of course, retain as a meaningful one: that between the *impossible* event ($E \equiv 0$, corresponding to the empty set), the *certain* event ($E \equiv 1$, corresponding to the entire set), and all the others (the *possible* events, which are structurally indistinguishable among themselves).

If one actually wishes to continue along these lines, by introducing new elements into the basis, or – if one prefers to put it this way – passing from $\mathscr{Q}$ to $\mathscr{Q}' = \mathscr{Q} \times \mathscr{Q}^*$ (the cartesian product with any suitable $\mathscr{Q}^*$) until each 'elementary case' Q of $\mathscr{Q}$ is subdivided into M elements, all the E (thought of as sets of $\mathscr{Q}'$), apart from the empty set, will have cardinality M (because the cardinality will be both $\geqslant M$ and $\leqslant M \cdot M = M$). In this way, only the empty set and its complement will be distinguishable from the others. It would, however, be cumbersome to actually reduce oneself to such a $\mathscr{Q}'$. We shall content ourselves, therefore, with the fact of having mentioned the possibility of this equalizing of the cardinality for all the events of interest, without insisting on it being done, or taking it into any further account. It is sufficient to state that we ignore as being irrelevant any distinctions made on the basis of considering cardinality. We shall not mention this again. In fact, without going into all the details, all this could have been regarded as implicit in the assertion that we were not going to acknowledge any distinction between the subdivisions of a partition Q, and the (corresponding) subdivisions of a finer partition $\mathscr{Q}'$.

Instead, we must go back to the problem of the nuisance structures introduced by the presence of the basis $\mathscr{B}$; the structures that we had decided to *ignore*. Rather than ignoring them, we can *make use* of them, by removing, in a different way, the drawback they had of inducing a special status for some events, or random quantities, in comparison with the others. Instead of prescribing that the basis be ignored (and let us suppose for the moment that we have a basis of events, $E_h \in \mathscr{B}$, $h \in H$), we can achieve the desired result by enriching the basis itself so that it includes all the $E \in \mathscr{E}$. It then follows that membership of the original basis $\mathscr{B}$ is no longer relevant.[10] In the case of a basis of random quantities ($X_{h \in} \mathscr{B}$, where $h \in H$), a thorough application of this same procedure takes us even further. More precisely, it will be a question of adding to the X_h of the basis $\mathscr{B}$ all the random quantities expressible as functions of them (*any* functions); that is every function X_k of the points $Q \in \mathscr{Q}$, $X_k = f_k(Q)$, where the $f_k(\cdot)$, $k \in K$, put every Q of $\mathscr{Q}$ into correspondence with a real number. (It goes without saying that we do not impose any restrictions like continuity, etc., because we have already said, and repeated, that we do not consider $\mathscr{Q}$ as having any structure, and so such restrictions do not even make sense.[11]) The field $\mathscr{S}$ will then be the Cartesian space – let us denote it by S_K – with an infinite number of coordinates $x_k (k \in K)$, where K is the set of the indices k that label the functions $f_k(\cdot)$ forming the field $\mathscr{S} = \mathfrak{c}^{\mathscr{Q}}$. Essentially, K is $\mathscr{S}$ itself (and it is only for

10 This is rather like everyone at birth receiving the title of 'Your Excellency' in order to achieve its downgrading (something which the abolition of the title would not, since there would always be a handful of people who would retain it).

11 The distinctions which clearly do make sense are those concerning the 'possible' values of the X_k (i.e. the range of $f_k(Q)$); the case of *bounded* X_k is particularly important (as we have often seen already in this work). Here we are not directly interested in this aspect, because it does not depend on things concerning the field $\mathscr{Q}$ (and related notions).

notational convenience that we introduce an index k, $k \in K$, in order to distinguish the functions $f_k(\cdot)$; it would be equivalent to refer to the functions $f, f \in \mathscr{S}$). One may observe that K (or, equivalently, S) has cardinal $\mathfrak{c}^M$, where M is the cardinal of $\mathscr{Q}$, and that K contains H, $H \subset K$, given that among the complete collection of $X_k = f_k(Q)$ there exist, in particular, the X_h which form the basis $\mathscr{B}$.

The 'waste' in the number of dimensions (in passing from H to K) is clearly considerable and might seem rather absurd. On the one hand, however, the fact that this enormous waste is more or less a disaster is irrelevant in practice, since it is never necessary, nor would it be possible, to take account of the infinite dimensions one at a time. On the other hand, such an extension brings with it something that is very much to our advantage (formally, for the time being, but of substance later, when we introduce into this framework the notions of probability and prevision). This advantage lies in the following: that, *by this means*, we could also *retain, and consider as valid, the linear structure* thus introduced into $\mathscr{S}$ and which makes it into a linear ambit $\mathscr{A}$, since – by the principle of 'everyone a nobleman' – it *no longer gives rise to any discrimination among the various random quantities and, in particular, among the events*.[12]

The extension made for $\mathscr{S}$ does not modify, in any basic respect, the set, or field, Q of possible points Q. Those that are (provisionally) considered as 'elementary possible cases' remain the same, but the 'points' representing them are dispersed and spread in the enlarged field $\mathscr{S}$ to a much greater extent. A basic intuition can be obtained from recalling the example of the parabola, $y = x^2$, given in Chapter 2, 2.8.7; however, this only concerns a single random quantity. In the general case (and also in the case of the example above, where one considers all the $Y = f(X)$, even restricting oneself to $Y = X^n$), it turns out that all the points Q (in the field $\mathscr{S}$ extended to a linear ambit A) are *linearly independent*. In other words, if $Q_1, Q_2, \ldots, Q_n$ are *possible* points – belonging to Q – then, in all the S_{n-1} which they determine, there is no other possible point (i.e. the intersection of $\mathscr{Q}$ with such S_{n-1} reduces to these n points, and in general consists of at most n points).

In order to fix ideas, let us verify this first of all for the simplest example, which we mentioned above. The field $\mathscr{S}$ is the Cartesian space with an infinite number of coordinates, x_h ($h = 1, 2, 3, \ldots$), on which are represented the values of the random quantities $X_h = T^h$ (where, for greater clarity, we denote by T the random quantity with which we begin; i.e. X_1, represented on the x_1-axis). The field $\mathscr{Q}$ is the 'line'[13] with parametric equations $x_1 = t, x_2 = t^2, x_3 = t^3, \ldots, x_h = t^h, \ldots$, if the random quantity T admits all the reals

12 In Bodiou (1964, *op. cit.*) we find further discussion of these topics, again in agreement with our views, and all the more interesting since his discussion is not inspired by abstract, conceptual questions like those we have raised here, but by problems in quantum mechanics. Arguing in favour of referring to a 'dialectic lattice' (like the one he proposes) rather than to a particular special form of Hilbert space, he makes the following remark (p. 103): 'The unwarranted special status conferred on the coordinates of the particle by this particularization obscures the general character of the notions, and gives rise to pseudo-problems (…) The coordinates are random quantities just like all the others, no matter how important they might seem.'

13 We use the word 'line' for convenience, it being a set of points depending on $t (-\infty < t < +\infty)$. For our purposes, it does not matter whether this term is really appropriate for some other aspects of the problem. (One thinks, for example, of the 'peculiar fact' that on the segments where $|t| > 1$, the points of the 'line', with the usual metric, all have infinite distance from one another. However, it would be sufficient to consider the modified line, $x_h = t^h/h!$, or, equivalently, to use the metric $[\sum_h (x_h / h!)^2]^{\frac{1}{2}}$, in order to overcome these difficulties. We mention all this merely for the sake of curiosity.)

(from $-\infty$ to $+\infty$) as possible values; otherwise, it is the subset of the points of the 'line' corresponding to the values $t \in I$ of the parameter t belonging to the set I of possible values for T (one should constantly refer back to the example in Chapter 2, 2.8.7).

To establish that linear independence exists between the points of $\mathscr{Q}$, it is sufficient to recall, for example, the fact that the Vandermonde determinant is nonzero. Given any n points of $\mathscr{Q}$ – $Q_1, Q_2, \ldots, Q_n$, say, corresponding to $t = t_1, t_2, \ldots, t_n$ – if they were linearly dependent (i.e. if they belonged to an S_m, $m < n - 1$) then, *a fortiori*, their projections onto an $(n - 1)$-dimensional subspace would also be linearly dependent. One could, for example, take the projections onto the subspace obtained by considering only the first $n - 1$ coordinates $x_1, x_2, \ldots, x_{n-1}$ (setting $x_h = 0$, $h \geqslant n$). This would imply the vanishing of the Vandermonde determinant $(a_{rs} = t_r^{s-1}; r, s = 1,2,\ldots,r)$, which is impossible for distinct values of t. It might be observed that the same proof also holds for all other infinite projections (but one is enough to establish the conclusion).

The proof of linear independence in the case of a general linear ambit A is even simpler: the preceding case is useful only in that it deals with a situation that is, in a certain sense, more immediate, because no discontinuities are involved (these arise in the consideration of events, with 0 and 1 as the only possible values). Let $Q_1, Q_2, \ldots, Q_n$ be once again points of $\mathscr{Q}$, and let Q_0 be a point which is linearly dependent on them:

$$Q_0 = a_1 Q_1 + a_2 Q_2 + \ldots + a_n Q_n$$

with $\Sigma a_h = 1$.[14] Let us divide the n points Q_h into two groups, labelling them with one or two dashes, respectively (i.e. writing Q'_h or Q''_h to indicate whether Q_h is in the first or second group). The only condition is that the sums a' and a'' of the weights a'_h and a''_h are neither 0 nor 1 (this can always be arranged, except in the case in which a single a_h is equal to 1, and all the others are zero; i.e. the case in which Q_0 coincides with one of the given points Q_h: we shall obviously exclude this case). We denote by E' and E'' the logical sums of the Q' and the Q'', and we let E be any event $E' \subset E \subset \tilde{E}''$ (in simple terms, we put the Q' in E, the Q'' in $\tilde{E}$, and we divide up all the other points of Q arbitrarily between E and $\tilde{E}$). For any such event E, we can say that it has the value 1 over all the points Q', and the value 0 over all the Q''; consequently, it has value a' on Q_0 (where $0 \neq a' \neq 1$). It follows that Q_0 is not a point of $\mathscr{Q}$ (and therefore not 'possible') because it does not attribute to E one of the two values 0 or 1.

It might well seem absurd to 'invent' – not without some effort – a field S, or a linear ambit $\mathscr{A}$, constituted almost entirely of points that satisfy ridiculous conditions (like making an event – whose values can only be 0 or 1 – assume values such as log 2, or π; or making X take on the value 1, and X^2 take on 2 or 0). It may be, however, that this makes sense in terms of probability and prevision (we might well have $\mathbf{P}(E) = \log 2$ or $\mathbf{P}(X) = 1$ with $\mathbf{P}(X^2) = 2$; this would happen, for example, if X took on the values 0 and 2, each with probability $\frac{1}{2}$), or in terms of linear combinations of previsions (see the footnote to Chapter 3, 3.7.2). Let us take advantage of this glimpse into the future in order to simplify our construction somewhat. Suppose that we assume (rather unjustifiably, because it runs a little ahead of the axiomatic treatment) that points satisfying

14 This notation really implies that Q_0 is the barycentre of the Q_h with masses a_h; i.e. the point whose barycentric coordinates, with reference to the basis-points Q_h, are a_h. In Cartesian coordinates, this implies that, for any x, the values $x^{(i)}$ which it assumes at Q_i $(i = 0, 1, 2, \ldots, n)$ satisfy a similar relationship, $x^{(0)} = a_1 x^{(1)} + a_2 x^{(2)} + \ldots + a_n x^{(n)}$.

linearly contradictory conditions are of no use (for example, $2X - Y = 3$, $X + 4Y = 1$, $3X - 6Y = 4$; we see that $3X - 6Y = 2(2X - Y) - (X + 4Y) = 2.3 - 1 = 6 - 1 = 5$). We can then restrict the linear ambits $\mathscr{A}$ to that part (of the larger construction) where things go through for linear relations.

As an alternative to this *a posteriori* exclusion of the superfluous part, one could simply avoid constructing it in the first place. Thinking in terms of a transfinite mode of construction, introducing one after the other the functions $X_k = f_k(Q)$, it is sufficient not to introduce new axes x_k whenever the X_k turns out to be a (finite) linear combination of those already considered. We shall make the convention that the (bogus) random quantity $X_0 \equiv 1$ be introduced as the first element (for the same reasons, and with the same effects, as in Chapter 2, 2.8.3) and we shall obtain the *linear space* $\mathscr{L}$ of the X, the dual of the linear ambit $\mathscr{A}$ (this follows in the same way as before, so we omit the details).

In order to illustrate all this in a more concrete fashion, it is convenient to refer to the construction we mentioned above. Some of the X_k (let us denote them by X_k^*) are represented on new coordinates x_k, whereas those linearly expressible in terms of a finite number of the preceding ones already find a coordinate available: $x = u_0 x_0 + u_{k_1} x_{k_1} + \ldots + u_{k_n} x_{k_n}$. This representation is one-to-one[15] and also holds for the preceding case (for $X_k^* : x_k = u_k x_k, u_k = 1$), which does not have to be regarded as special in any way. Each point Q, and similarly each point A of $\mathscr{A}$ (even those not possible), is characterized by the values $x_k(Q)$, respectively, $x_k(A)$, of its coordinates on the axes of the X_k^*. For every other X, the value will be given by the above finite linear combinations (of the coordinates x_{k_i}, calculated at Q or at A, respectively). These values could be written (just as in Chapter 2, 2.8.3) as $A(X)$ or $X(A)$, and interpreted as products of vectors (from $\mathscr{L}$ and from $\mathscr{A}$), with the sole difference that instead of linear combinations from among a finite number of elements $(n + 1)$, we have finite linear combinations from among an infinite number of elements.

We shall call a halt here to our description of the formal set-up. The consideration of these topics – although necessary for a complete presentation of the scheme – has already led us so far into the probabilistic meaning that we cannot usefully say anything more without bringing in the latter explicitly.

4 Questions Concerning 'Possibility'

The most clear-cut distinction in the entire formulation is that between possible and impossible events (or certain ones, but these are merely complements of the impossible ones). But is this distinction really so clear-cut?

There would seem little room for doubt. Someone who argues that he does not know whether an event is possible or not is, in actual fact, already saying that for him it is possible (because he cannot exclude it, as would be necessary if it were impossible). In the same way, when one says that 'it is not known whether, in a census, the sex of a particular individual is known or not', one is already admitting that it is 'unknown'. As a matter of terminology, this is without doubt absolutely correct. When we try to apply

15 Were it not so, there would exist a linear relation among a finite number of elements X_k^*, and then the *last* one would have been mistaken for an X^*.

the principle to actual situations, however, and we start examining the exact nature of the dichotomy between what one *knows* and what one does *not know*, the dividing line seems much less absolute and various kinds of difficulties arise.

The most serious of these is the one concealed in the very logical mechanisms which exist to overcome it. We know that every consequence of something which is certain is itself certain; in other words, all that is implicit in those things for which we have explicit information must be considered as part of that information. It follows that the field of what is certain for someone – that is for which information is available – must be *closed* with respect to deduction; in other words, it must not leave outside anything which is deducible. But deducible in what sense? Through the mechanisms of logic, and this, as Galileo says in a celebrated passage, requires 'voyages of our mind, step by step, with time and with motion'; but for things to be as easily done as said, we should require 'the mind of the Almighty', voyaging 'with the speed of light'.

The extreme case, in an opposite sense, is that in which, for some individual, certain conclusions are completely beyond reach; either because his knowledge is insufficient, or because he is not capable of the necessary reasoning. Suppose that N is the number of paving stones forming some given rectangular pavement (for example, one that the individual recalls having seen, but for which he has only a vague recollection of the dimensions, and of the dimensions of the paving stones). The set of possible values for N should, at the very least, exclude the prime numbers. But what if he is not familiar with this notion? Or if he was familiar with it once upon a time, but finds that things he learnt at school do not come back to him sufficiently readily when he is faced with problems where such knowledge would be useful? In these circumstances, he could not even contemplate making such an exclusion. This extreme case is in any case the simplest; the set of things that are 'certain' always remains closed with respect to this deductive capability (even if in an incoherent fashion, since it does not coincide with the logical possibilities).

The situation is more awkward when 'closure' fails to hold and is replaced by something less rigid. The clear-cut distinction between those certainties that an individual can work out for himself and those which he cannot is lost, and instead we have certainties that are attainable with various degrees of difficulty, and perhaps not immediately. In terms of the above example, this is the situation faced by an individual who knows that he has to exclude the prime numbers, but who finds that for many integers it is not easy to see at a glance whether or not they are prime. Is it worthwhile making all the calculations required in order to find out? Or is it worthwhile checking through a table of prime numbers (first searching for such a table and then searching for the number in question)? It would probably not be worthwhile if the purpose were merely to exclude certain rare numbers from the set of those to be considered as 'possible'. It might be the case, however, that, by virtue of some additional information, we were in a position to determine N uniquely, since it would turn out to be the only possible value. (As an example, suppose we knew that the diagonal of the rectangle were a multiple of the side of the paving stones. It follows that $N = XY$, where the sum of the squares of X and Y is a square; if we had sufficiently close bounds, $x' \leqslant X \leqslant x''$, $y' \leqslant Y \leqslant y''$, or some other similar information, X, Y, N would turn out to be uniquely determined.) In such cases 'is it worthwhile'?

In order to answer this question, we need, of course, to know what the alternatives are, and what degree of interest they hold for the individual relative to his evaluation of

the difficulty and the labour of finding out. Among the possibilities, we note the following: we are not interested in the problem, not even out of curiosity, and we have the option of not bothering ourselves with it; we are content with a crude estimation (for example, we may wish to buy n paving stones, choosing n such that we are almost certain of there being sufficient to enable us to form a pavement of the same size as the original, but with not more than 10% of the paving stones left over; $\mathbf{P}(N > n) = 1\%$, $\mathbf{P}(N < 0{\cdot}9n) \leqslant 5\%$, say); if, instead, there is a lot at stake, we can make precise our distribution of probability, $p_h = \mathbf{P}(N = h)$, and compare the costs, risks and advantages of buying this or that number of paving stones.

The aspect that concerns us most here is the difficulty of verifying just what it is that is implicit in our initial data. Sometimes this difficulty will be insurmountable (evaluate the billionth decimal place of π; check the Goldbach conjecture up to 10^{100} etc.; and it would be even worse were we to consider things of this kind involving an infinite number of digits or integers); at other times, the difficulty is simply one of the labour or cost involved. We shall encounter these matters again in connection with the verifiability of events, and we shall then go into them more deeply. At this juncture, it is sufficient to observe that, from a practical point of view, there is, in the last analysis, no difference between an experiment or investigation aimed at uncovering information about an unknown fact, and an attempt at deduction aimed at ascertaining that which, in theory, we should already know on the basis of the information in our possession. The difference is simply that if we abandon such attempts at deduction (because of more or less insurmountable difficulties), the field of that which we 'actually know' is incoherent, since it is not closed from a logical point of view.

We obtain maximum 'flexibility' if it is easy to demarcate the field of those cases 'possible at first sight'; the difficulty here, however, is that it takes a long time to check all cases. (To give a simple, though rather silly, example: all the integers n between 1 and 1000 excluding those for which 5 appears at least twice in the first six decimal places of the reciprocal.) To check through to the very end presents difficulties, but the same is true of stopping at some arbitrary point and, therefore, of starting off with any rigorously laid down distinction between those cases which are possible and those which are not.

Such a distinction becomes even weaker if we abandon the convenient hypothesis that the knowledge one starts with is precisely specified in the stated 'data of the problem'. If we trace back to the actual, empirical source of the knowledge, are we really able to draw a reasonable line between possible and impossible? If a person had seen our pavement he might very well say that he is 'certain' (but is it true?) that it could neither be less than $1 \times 2\,\mathrm{m}^2$, nor greater than $100 \times 1000\,\mathrm{m}^2$ (suppose he estimated it as about 4×7). But can such an absolute logical distinction be declared valid in one case and yet one not know at what point it ceases to be valid? Is the length of 3 m too small to be possible? Yes? And a length of 4 m? No? Then what about the limit being 3·42 or 3·423 m? Can it not be specified precisely?

It would perhaps be appropriate to only use the phrase 'absolutely certain' when referring to tautologies (and even then …?). Adopting this rule in our example, we could only say that the side may have any length between 0 and ∞. In fact, nothing prevents us, either for convenience or by convention, from restricting our consideration of the problem to a narrower and more 'reasonable' set of values (like those mentioned above, 1×2 and 100×1000, or to an even more restricted set). It would, however, be more prudent and responsible to substitute in place of 'it is certain that …', the phrase 'assuming

that … (the risk of my being wrong appears to me completely negligible)'. Can we talk of an age limit beyond which it is certain that some given individual (or any person living at the present time) cannot survive? Or of a speed which cannot be exceeded (for example, in a track race, or swimming 1000 m, or by a bicycle, car, aeroplane etc.)? I would say that it is always inadvisable to express in terms of 'certainty' any everyday judgement of this kind (even on those occasions when I share the judgement, though with a less firm conviction).

So what conclusion do we come to? We could well repeat the statement that we made initially: that if we do not know whether something is possible or impossible, then, by definition, it is possible. Let us bear in mind, however, that everything is based on distinctions that are themselves uncertain and vague, and which we conventionally translate into terms of certainty only because of the logical formulation. On the other hand, even in the case of those in a census for whom 'it is not known whether or not the sex is known', it may very well be that the doubt does not have to be resolved straightforwardly by saying 'therefore it is unknown'. This is true, certainly, but if we are considering the possibility of further investigation, allowing – either definitely or potentially – the completion of the missing information, the nature of the problem changes.

In the mathematical formulation of any problem it is necessary to base oneself on some appropriate idealizations and simplifications. This is, however, a disadvantage; it is a distorting factor which one should always try to keep in check, and to approach circumspectly. It is unfortunate that the reverse often happens. One loses sight of the original nature of the problem, falls in love with the idealization, and then blames reality for not conforming to it.

In our case, it is certainly necessary that we base ourselves, in the initial formulation, on the distinction between possible and impossible; the distinction being considered as clear-cut as suits our purpose. We must be on our guard, however, not to become prisoners of this artificial rigidity (of absolute certainty) if, rather than helping us, it should come to trap us in an incomplete and distorted view of things. We shall return to these questions, and delve more deeply into them, both in relation to impossibility (as above) and with (new) reference to zero probability, and, above all, in relation to the acquisition of information for the purpose of an evaluation of probability. It is only then that we may come to have a more precise understanding of the questions we have had to illustrate here in terms of one particular aspect, without considering the complementary one.

5 Verifiability and the Time Factor

We can admit that an event E, suitably described, makes sense objectively; that is, is true or false independently of any possibility that we have of knowing the fact. To affirm or deny its truth as a general thesis would mean that one was being metaphysical, but it cannot be doubted that it is often convenient and almost unavoidable to think in this way (albeit with due caution).

However, what is important for our purposes is not the fact that E is objectively true or false (if one can speak of a 'fact' in this connection). What really matters is to establish the fact (to obtain the information, to verify) that E is true or E is false. Think in terms of a bet involving E: this is the most futile, but nonetheless the most expressive, example

for demonstrating that what counts is only the knowledge of the outcome. Moreover, the case of a bet serves as a typical model of the general situation in which probabilities serve as guidelines for decision making under uncertainty and for applications within this framework.

Let us return now to the study of the field of possibility, which we began in Section 3 (without restrictions; and let us leave aside, for the time being, the reservations made in Section 4). Let us assume that in this field every event is – in an 'objective' sense – either true or false (and that every random quantity X has a precisely determined value, x); in other words, we assume that one may imagine as uniquely defined that point Q of $\mathscr{Q}$ which summarizes the truth or falsity of all the events of $\mathscr{E}$ and the exact values of all the random quantities of $\mathscr{X}$.

It might be thought that such a schematization is excessively theoretical and pretentious, even for the representation of the situation relating to a theoretical conception like that of 'objective' reality (which does not take into account limitations deriving from imperfections of ourselves and of the instruments with which we make our observations). It might even be said (from the opposite point of view) that it does not make sense, because things only acquire meaning through, and as a result of, observation and measurement. Either way, it seems that one is less open to criticism in starting from such an 'overdone' schematization of some 'objective reality' (that one may or may not take seriously), since this is merely the starting point for the introduction, as and when it suits us, of the gradual qualifications that take place in the transition from the metaphysical notion of 'objective truth' to the effective notion of 'verifiability'.

As we have already mentioned in Sections 1 and 2, these modifications and qualifications will have to be examined from various points of view, in relation to various circumstances and factors. As a first step, let us consider the *time* factor. In order to simplify the question, and to separate it off from others that are almost always connected with it, we restrict ourselves to the case in which, from a given instant t onwards, the result is known to everyone (or, at any rate, accessible, there being no need to do anything in order to ensure its occurrence or to learn about it). In order to fix ideas, think in terms of the entire output of news from the press, radio and television (political news, day to day items, weather, sport, the economy, science, the arts and so on).

Several cases may arise for each event E, according to the instant at which the result is known. An event may be *dated* more or less precisely, in the sense that the instant at which one knows whether it is true or false is known *a priori* (or, if we are dealing with a random quantity, the instant at which its value is known). For instance, the maximum temperature in Rome in August of next year (together with the fact of whether or not it is greater than that of the preceding August) will be known immediately after the end of the month in question. The population of Italy at the next census (the year of this being fixed by law, the actual day not yet decided) will be known shortly after the time mentioned (i.e. the year already fixed by law and the day chosen); and the same is true for the question of whether the population of the North or of the Central-South has had the greatest increase. It is easy to think of other examples.

In other cases, there exists a maximum time limit (possibly rigid, possibly not) before one knows the answer; and this advance knowledge could be relevant to only one form of answer (either affirmative or negative), or to both. As examples, consider the fact of an individual remaining alive, or continuing to hold his present post without interruption, or never being sick, or never having a car accident, and so on, up to some

preassigned time or age. These are all statements that can only be ascertained as true at the last-mentioned time (if at all; for it could turn out that they were known to be false at some earlier instant). As for examples the other way around, it suffices to consider the negations of the above statements. If the preassigned time limit is dropped, and one stipulates 'until death', the conclusion[16] remains the same, except that a maximum time limit can no longer be given with certainty (although in practice the time at which a person would become 100 years old might be considered appropriate for the purpose in hand). The asymmetry between affirmative and negative answers vanishes if, always with respect to the life of a given individual, we consider examples like his dying because of an accident, or some other cause, or before (or after) some other individual, or in Italy, or abroad. (If in these examples one wishes to include a time limit, it is necessary to decide in advance which result is considered as valid if the limit is reached; for example, if the individual were insured against one of the two eventualities up to the time limit, to reach this limit would be equivalent to the occurrence of the other.)

There are cases, however, in which a statement can never be either verified or disproved until the end of time (or can only be verified or only disproved, or the one and the other, but it is not known if and when). Obvious examples of statements that can never be settled one way or the other within a finite time period are easily found; one only has to consider sequences of events unbounded in time. For example: tosses of a coin; spins of a roulette wheel; rainy and dry days (in a given locality); normal days and those days on which men turn into rhinoceroses;[17] male and female births (in chronological order, in some given town); passages of people in one direction or another through a gateway; and so on. Taking the coin tossing example, for convenience, it is sufficient to say, for example, 'neither Heads nor Tails will always occur', 'the frequency of Heads will tend to $\frac{1}{2}$', or 'will tend to some limit', or 'will exceed both the bounds 0·01 and 0·99 infinitely often', 'from some point on Heads and Tails will alternate in a regular fashion, HTHTHT…', 'the sequence HTTHH…, which represents the Divine Comedy in binary code, will be repeated an infinite number of times', and it would be easy to go on in this way. If instead we were to say that 'the frequency will be less than 0·01 at least once (after the first 1000 trials)', or 'we will have (after the first 1000 trials) at least one run of Heads as long as the preceding segment' (i.e., starting at the $(n + 1)$th toss it reaches at least to the $2n$th), and so on, we would have examples of statements which, if true, are certainly verifiable within a finite period of time (although we do not know how long). For the complementary statements the opposite holds. If we add to each of the above statements something like '… or the frequency tends to $\frac{1}{2}$', we have statements which, if true, can either turn out to be verifiable within a finite time period, or not. An example of a statement that may or may not be verified within a finite time both if true and if false is the following: 'if one has, at least once (after the first 1000 trials), a sequence of identical results (always H or always T) from toss $n + 1$ to toss $2n$, this will occur for the first time for the outcome Heads; if this never occurs, the frequency of Heads will have lower limit $\geqslant \frac{1}{2}$'.

16 Except for the first example, which becomes meaningless.

17 Up until now, all days have been 'normal' (Ionesco notwithstanding), and I think that this will always be the case. An example of this kind was needed, however, in order to make clear that it in no way differs from the others insofar as verifiability is concerned.

All the events we have considered, expressed as statements about some given sequence of 'trials' of unlimited duration in time, or concerning phenomena (like the death of an individual) that can take place at various times, can be more completely described by means of the ordered pair (E, T) (event E, taking values 0 and 1, and random time T, with $0 \leqslant T < \infty$, or $T = \infty$). This enables one to identify not only the truth value (True or False) but also the time (finite or infinite) after which the event turns out to be verified. A useful convention, enabling one to reduce the ordered pair to a single random quantity, is that of taking T with a + or − sign, according to whether E is true or false: that is putting $T^* = T.(E - \check{E}) = T.(2E - 1)$ (which, in fact, gives $+T$ if the time is T and $E = 1$ = true, $2E - 1 = 1$, and gives $-T$ if the time is T and $E = 0$ = false, $2E - 1 = -1$). We can summarize the various cases by expressing them in terms of this random quantity T^*. Both for positive values and for negative values (and independently of what happens in the other case) the possible values for T may either reduce to the unique value ∞, or to ∞ together with some finite values, or to finite values only (and in this case either unbounded or bounded; possibly 'bounded in practice', to put it in an unorthodox way).

What happens if we consider logical operations on events whose verification may be put off until different times, possibly 'never'? It is clear that under negation T^* changes sign; if $E_2 = \tilde{E}_1$, $T_2^* = -T_1^*$. If we consider the (logical) product, $E = E_1E_2$, E must turn out to be true if and when this has happened both for E_1 (at time T_1) and for E_2 (at time T_2), that is at time $T_1 \vee T_2$ (the larger of the two). In the opposite case, E will turn out to be false as soon as either E_1 or E_2 does; that is either at time T_1 or at time T_2, or at the smaller of the two if both events turn out to be false. All this can be condensed by means of the following convention: for $E = E_1E_2$ we have $T^* = T_1^* \vee T_2^*$ with the convention of modifying the general meaning of the sign as follows; 'take the larger of the negative values (the smaller in absolute value) and if they are both positive take the larger'. It is easy to see that this rule also holds if we include the values $+\infty$ and $-\infty$, and that it can be extended to deal with the product of an arbitrary number of events. By means of this rule, one can construct the set of possible values of T^* starting from those for T_1^* and T_2^* (and possibly others) *provided they are logically independent.*[18] Moreover, given that the logical sum is the negation of the product of negations, it is immediate, from what we have said above, that for $E = E_1 \vee E_2$ (and also for several events) we have $T^* = T_1^* \wedge T_2^*$, with a convention dual to the previous one: 'take the minimum of the positive values, and if they are all negative, again the minimum (i.e. the maximum in absolute value)'. On the other hand, these rules are obvious if we think of the meaning of the actual problems themselves.[19]

Actually, it is clear (and it also follows formally from what we have said) that logical combinations of events which are certainly verifiable within finite time periods yield events that are (in general) certainly verifiable within the most extensive of these time periods.

---

18 The meaning is the usual one: however, it might be useful to discuss it in our present context. T_1 and T_2 are logically independent if, when t_1' is a possible instant for E_1 turning out to be true, and t_1'' is a possible instant for it turning out to be false, and the same is true for t_2', t_2'' with respect to E_2, it is also possible that E_1 turns out to be true at t_1', and E_2 false at t_2'' (and similarly, interchanging true and false, for t_1'' and t_2', t_1' and t_2', t_1' and t_2').Of course, if logical independence did not hold, the set of possible values for T^* would either still be the one so determined, or a subset of it.

19 In order to avoid confusion, it would certainly be useful to introduce modified notation to replace $\vee$ and $\wedge$ if we planned on making actual use of it. In fact, we are only going to use the idea here, temporarily, for the purpose of this explanation: it is, therefore, not worth complicating things.

If, however, we consider an infinite number of events, which are each certainly verifiable within finite, although unbounded, time periods, we no longer have certain verifiability.

Let us now turn to the examination of the points raised by our analysis of the circumstances surrounding the notion of an event (with special emphasis on the realizability of a bet, which serves for us as something of a 'touchstone'). It seems natural to conclude that every postponement, and every asymmetric feature (between the ascertainment of the truth of either result) which might be caused by it, has an adverse effect on those characteristics required for an event until, if the postponement is too great, or even forever, these characteristics completely disappear.

Although we have to postpone the most relevant comments until after the introduction of probability, it is certainly clear, even at this point, that it would be very strange to discuss the truth of a statement, or bet on it (or even to maintain that it makes sense), when it does not assert anything that would enable one to discriminate in any way between possible future observations, even were one to think in terms of living for ever, or of passing on the task to future generations (assuming they never become extinct). And how silly it would be, besides being strange, to bet on a statement constructed in such a way that it is possible to lose right now, but to win only after the end of time-without-end (and this is so, even if it is a statement involving only a very small risk – like the assertion that 'the day on which men will turn into rhinoceroses will never come').

6 Verifiability and the Operational Factor

In Section 5, for the purpose of isolating the *time* factor, we restricted ourselves to considering the result as 'known to everyone (or, at any rate, accessible, there being no need to do anything in order to ensure its occurrence, or to learn about it)'. This assumption is not tenable as anything other than an idealized limit case, because even to listen to the radio, or to read a newspaper, requires some time, effort and cost, even if only a small amount. In general, however, it is necessary to do a great deal more in order to check the truth or falsity of an event or statement. It is often necessary to actually experiment in order to observe, or measure, or even produce, the phenomenon under consideration. In any case, even the mere recollection of existing data, or the task of researching for information concerning data already collected, can be involved, non-trivial operations.

We use the term *operational* factor to describe anything which, by virtue of the nature of such operations, imposes constraints on the verifiability of events. One aspect of this is the *cost* factor (intended in a broad sense), which we shall mention in conjunction with it. 'Precision' and 'indeterminacy' are other factors closely connected with the operational factor, but, because of their importance, we shall treat them separately in Sections 7 and 8.

In the first place, we come across a feature similar to that encountered when we dealt with the time factor: this is the difficulty, or even impossibility, of performing an excessive (or infinite) 'number' of operations (we say 'number' even though the terminology only makes sense in certain cases). For example, if we want to ensure that in a given time interval, $t_1 \leqslant t \leqslant t_2$, a certain quantity $y = f(t)$ has not exceeded a given level $y = y_0$, and for this purpose we wish to measure y (or just to check that $y \leqslant y_0$) at *each* of the

infinitely many instants of the interval, the task would appear impossible to carry out. (And the same would be true even if measurements were only made at a dense subset of time points – also infinite, even if denumerable.) This statement is not, of course, to be taken as indisputable, deriving from some metaphysical prejudice, but simply as an empirical observation that seems undeniable in many practical instances (and in cases where it could be challenged we will acknowledge the fact). One cannot, however, conclude from this that there is no way of verifying the event under consideration. For the example in question, it suffices to invent and install some device like a 'max–min' thermometer, as used for temperatures, or like a fuse, calibrated to take an upper limit of electric current.

In the example we have considered, note the following two circumstances. Firstly, the constraints thaat derive from the impracticability of simultaneously considering an infinite number of trials might, formally, be the same as those of Section 5, but the significance is entirely different. In this context, it is not the *postponement* of the verification, *sine die*, which we are concerned with, but rather its *unrealizability* (were one able to do it, it could be done in a finite time). Secondly, the fact that a statement is not verifiable by means of some given procedure (here, measurements at an infinite number of instants) does not preclude its being verifiable in some other way. Unverifiability in some absolute sense cannot be asserted on the basis of the unrealizability of some or all of the operational schemes put forward so far: it can only be asserted on the basis of some rather general assumption that excludes realizability under *any* scheme (and the basis for such an assumption may or may not be very secure …).

A formulation that might in certain cases adequately express the meaning of such constraints (albeit in a very schematized and idealized manner) could take the form of considering an event – or, better, a partition into events – as verifiable if one could reach it by means of a finite number of elementary, realizable operations. Note that if one regards the results of these operations as basic events this corresponds to assigning some special status to them – the very thing we strived so hard to eliminate! There is, in fact, no contradiction. In the first place, we have here made precise which criterion, if any, is to determine the basic subdivisions, and, hence, to assign them special status. Secondly, this would be justified only if in certain cases a formulation like the one we have just put forward as a hypothetical example appears to be actually valid (or, at least, almost so).

A similar, but more realistic, limitation (and not only from a practical point of view) would consist in restricting ourselves not simply to a finite, though arbitrarily large, number of operations but to a number not exceeding some given finite upper bound. In the case of the measurement of $y = f(t)$ in (t_1, t_2), for example, we must not merely take it to be impossible to make an infinite number of observations within the time interval, but also impossible to make more than some given finite number, which can be specified more or less precisely. In actual fact, this upper bound will, in general, be anything but precise unless we introduce the *cost* factor. Usually, one does not find a clear-cut point of separation up to which we can proceed without any difficulty but beyond which it is impossible to proceed. On the contrary, the fact of the matter is that one encounters ever increasing difficulty as one proceeds further and further. We use the term *cost* to denote the measure of these difficulties. As we have already remarked, it is a question of 'cost in a broad sense'; not simply the money spent but also the efforts made and the time required, taking into account the other alternative uses to which the time and effort might have been devoted.

We shall, however, always express this *cost* in monetary terms. This is done not so that we can adopt the economist's approach to the problem but simply to enable us to note that the problem of limitation can then be put in the following terms:

- First version: given the total budget available (over a certain time period), one can work out whether a certain sequence of operations is realizable or not. One must then select the most efficient from among those that are realizable (i.e. the one which gives the best overall result).
- Second version: this is a refinement of the first one and, following along the same lines, assumes that the best realizable sequence of operations is determined for a given total budget. The latter is no longer regarded as given and fixed, however, but as variable within some given range of values. In this version, the cost is also a matter of choice and will be chosen in such a way that one arrives at the equilibrium point in the neighbourhood of which an increase in cost produces an equivalent increase in efficiency (the principle of marginal returns).

The simplest assumption, so far as costs are concerned, is that of additivity (a given cost for each separate operation); in general, however, it is not necessary to limit oneself to this special case.

A further step towards realism consists in taking into account at least one particular kind of uncertainty which, in general, affects the 'operations' we employ in order to verify events (this is in fact the final such step we shall take: it seems to be adequate for the purpose of an idealized schematization). So far, we have assumed that such operations should always give us a precise answer, either YES or NO, to the question posed. We now make the assumption that the answer could also be MAYBE; in other words, either the experiment does not succeed, or it gives a result which is not sufficiently clear-cut to enable us to consider either YES or NO as established beyond doubt.

It is commonplace to remark that this can happen in any kind of experiment or procedure (such as the examination of a witness's testimony in court). If there is any well-founded objection to this statement, it would be that admitting MAYBE does not go far enough and that a clear-cut and definitive YES or NO cannot be obtained from any experiment … and that it is always a question of greater or lesser probabilities. It goes without saying that I very much agree with this but, in order to avoid getting into a vicious circle, it seems to me that the best one can do, or, at any rate, the least objectionable alternative, is to go along with this approach. It introduces uncertainty in a meaningful way, by including the possibility of the answer MAYBE, but does not preclude the construction of a scheme coming before the introduction of the notion of probability. Such a preclusion would arise if the woodworm of uncertainty were to find its way into the answers YES and NO. On the other hand, a study based on a formulation that was completely rooted in uncertainty (that is to say probability) could be carried out once probability theory was constructed. This is done, for example, in the classical theory of errors, where an error can be arbitrarily large, although with an extremely small probability (given by the normal, or Gaussian, distribution).

With respect to an event E, an operation may, therefore, yield either the answer YES (and hence NO for $\tilde{E}$), or the answer NO (and hence YES for $\tilde{E}$), or the answer MAYBE (the same for both E and $\tilde{E}$). But it may very well happen – and this is the case we shall consider – that the operation yields similar answers for other events, and ultimately for a partition, $C_1, C_2, \ldots, C_s$. In this case, the following possibilities arise: either the answer

YES is given for one of the constituents C_k (and hence NO for all the others – a complete answer); or there are no YES answers and more than one MAYBE, with or without NO answers (all MAYBE answers give an absolute void, leaving one in the same state of ignorance as before; if there is at least one NO we have a partial answer). With respect to the event E under consideration, the answer will be YES if E contains the constituent with the YES answer, or (if YES was missing) if it contains all the constituents with the answer MAYBE; the answer will be NO if, symmetrically, E is contained in the union of the constituents with NO answers; it will be MAYBE otherwise (i.e. if both E and $\tilde{E}$ are compatible with the union of the NO constituents and with that of the MAYBE constituents, replaced by the YES constituent if it exists).

Taking account of the partial answers in this more detailed manner – that is referring to the partition into constituents – greatly increases the efficiency of the procedure, because it permits us to draw the maximum possible information from the result of each operation. As an intuitive illustration of this obvious mechanism, suppose that in picking out the guilty party from among n individuals (who constitute the entire collection of possible suspects) we obtain $n - 1$ observations having the very limited effect of excluding one with the answer NO, and attributing MAYBE to all the others. If each single observation excludes a different individual, the unique remaining person must be the guilty party; if instead one had been concentrating on the guilty party (perhaps because he was the main suspect) and had only recorded whether or not each trial was sufficient, one would have had to conclude MAYBE, having always obtained the answer MAYBE.

So far as the applications to verifiability are concerned, this enrichment of the possibilities makes life somewhat more complicated, but compensates for this by removing some of the other complications deriving from the rigidity of the previous scheme. We shall see this particularly when we come to deal with measurement procedures (in Section 7).

On the other hand, everything could be expressed in a more direct and straightforward way by simply referring to the field Q of elementary cases Q (rather than fixing one's ideas – as, in a certain sense, is more instructive – on that subject which the operation has brought into play). Thinking in terms of Q, an 'operation' has the intended purpose of obtaining information, which means, in fact, narrowing the field Q by eliminating the Q that have turned out to be impossible (NO), and retaining the others (MAYBE). The YES would only be of use in the case of complete information, in order to pin-point the unique nonexcluded 'point' (or never, if one takes strict account of the observation that no subdivision – even if words like point or atom are used – can ever be considered as definitive, or as the ultimate one).

So far as the previous considerations about the advantages deriving from efficiency and cost are concerned, the only difference lies in the fact that the factor of uncertainty is introduced. If we decide to proceed in some given manner, it is no longer known *a priori* whether, and after how many operations, we will reach the desired conclusion. To judge a procedure as appropriate therefore necessitates the application of the theory of decision making under conditions of uncertainty – that is the maximization of expected utility corresponding to an uncertain cost and uncertain efficiency.

The experimenter could, taking everything into account, estimate, on a rough and ready basis, the procedure that he thinks to be the best. If he decides to apply the calculus of probability, so much the better (provided it is worthwhile to do so; i.e. that the

additional cost of performing the calculations does not exceed the expected increment in utility for choosing the optimal procedure). In any case, in this context we are not drawn into the vicious circle we previously indicated that we wished to avoid. The possible application of the theory of probability to this aspect of the problem is something which concerns the experimenter; he, independently of the stage in the present treatment at which we have occasion to speak of him, might equally well know or not know the calculus of probability. On the other hand, the fact of whether he has worked things out for himself well or badly (i.e. of whether he chooses the procedures in a more or less advantageous manner) is something of no concern or interest to us. Here, we are only interested in procedures in the abstract, in principle, as instruments we can make use of, and which give certain types of answers. Whether they are used well or badly is a separate question.

7 Verifiability and the Precision Factor

We have already (in Section 6) made some mention of measurements but, in order to divert attention from the topic being considered, we tacitly assumed that we were dealing with exact measurements. It is well known, however, that exactness is unattainable (except in counting procedures – provided one makes no mistakes). When dealing with measurements, one can only proceed by fixing in advance some higher or lower level of accuracy or *precision*.[20] Here, too (as in the case of similar questions considered previously), an improvement in precision generally implies an increase in cost. Looked at from this point of view, there is nothing new, and nothing to add to the above.

The most important issue, which we must examine, concerns the implications of imperfect precision in the measurements for the identification of the individual points Q of $\mathscr{Q}$ (or of $\mathscr{S}$). This leads into a discussion of whether, and in what sense, it is appropriate to introduce a topological structure into the field $\mathscr{Q}$ (or $\mathscr{S}$). We have, of course, struggled hard to eliminate any trace of such structure – as in the case of the special status of basis events – but here, as we found in that case, too, there is no contradiction. It will simply be a question of rejecting structures that are either unnecessary or suggested by so-called motives of analytic convenience and possibly accepting, after careful examination, those structures that correspond to essential and meaningful requirements.

Let us begin by considering the case of a single random quantity X, having the real line as its set of possible values. Thus far, in beginning our discussion of 'verifiability', we have considered events rather than random quantities. It is clear, however, that 'to verify' the value x taken by X is (precisely) to verify which of the events $E_x = (X = x)$ is the *true* one: in a weaker sense, it would be a question of verifying events of the form $E_I = (X \in I)$, with I, *a priori*, arbitrary.

There is nothing to be said on a general level about the possible ways of proceeding. This is not a mathematical problem, but rather one of the peculiarities inherent in each individual case. There is no problem when X appears already in a simple form, written in Arabic numerals (like the number on the ball at bingo, or on the sectors of the roulette wheel), or spelt out in the form of dots (as on a die), or is provided and vouched for by

20 In probability theory (and especially in relation to 'error' theory) the term 'precision' is often used to denote the reciprocal of the standard deviation, $1/\sigma$. Here, however, we use the word informally, without reference to any 'technical' meaning (as in Chapter 12, 12.3.1).

others (like the data from a census, or statistical data in general). In these cases we are dealing with integer random quantities and this, of course, is a simpler situation. It is possible, however, to find similar, immediate and rather precise representations when dealing with continuous quantities (physical constants, geographical coordinates, geodetic points, heights of buildings, weights of objects and individuals etc.). If, on the other hand, it is our responsibility to pin-point (exactly, or in part, or approximately) the value x of X, we have to examine, case by case, which events E_i are more or less accessible to us, and which of them yield the more interesting or useful pieces of information about X. Having done this, we then choose (as in the case already noted in Section 6) that combination of operations which is most advantageous, taking into account considerations of both efficiency and cost.

In what way is it possible to learn something about a random quantity X? And what can we learn? Posed in these terms, the question does not make sense because the answer does not depend on the fact of X being a random quantity, but rather, case by case, on the actual and particular meaning attached to each given X by virtue of its definition. The case usually dealt with, to the virtual exclusion of all others, is that of a physical quantity for which one may obtain more or less precise measurements; this will also be our principal concern here. It should be noted, however, that, apart from this rather special case, there is no reason to think that, in general, the problem can be posed in terms of the same concepts. An X defined as a (practically speaking) insufficiently continuous[21] function of another random quantity (having physical meaning) clearly does not lend itself to measurement through such a definition; but it may be measurable by virtue of the fact that it, or some suitable function of it, has a physical meaning of its own, which renders it capable of direct measurement. On the other hand, it may happen, especially if the definition of X is bound up with mathematical concepts, that the question of whether X belongs to sets I which are less straightforward (and in general less resolvable) than the intervals turns out to be easily answerable, or, at any rate, feasible (e.g. one might ask whether or not X is rational, or algebraic etc.).

The random quantity $X = \pi$ provides a suitable example. This might seem rather strange because – it might be argued – π is *not* a random quantity; it is a well-determined number, already determined with remarkable accuracy (given his time) by Archimedes, and for which explicit expressions in the form of series have been discovered, so that it is now known 'more precisely than any other number'.[22] Agreed: but π is not known (and is thus random, since we do not accept a more restrictive use of the term) insofar as its remaining decimal places are concerned; just as, for a long time, it was not known whether it was rational or algebraic. These questions, as everyone knows, have now been resolved (negatively): it is clear that this would not have been achieved merely by the more and more precise determination of the numerical value of π – even by going on to infinitely many decimal places – in order to look for possible periodicities (supposing it were rational) and so on. Instead, there was another approach (just as in the previous example concerning the maximum of a function in an interval).

Another example (of a simpler but more artificial kind) is provided by the following. Suppose we define $X = (N + 1)(N < \infty) - 1 + E'/4 + E''/\pi$, where π is no longer considered as a random quantity, E' and E'' are arbitrary events (e.g. a sports result and some feature

21 For example, in the sense of there being a Lipschitz condition, $|f(\xi_2) - f(\xi_1)| < K|\xi_2 - \xi_1|$, with K not too large (at least in the range of practical interest).

22 To 100 000 decimal places; given by D. Shanks and J. W. Wrench in *Math. of Computation* (1962).

of the weather), and N is the number of (possible) exceptions to Goldbach's conjecture.[23] The first two terms in the expression for X simply denote N, to be replaced by -1 if it is infinite; it is, therefore, an integer if the last two terms are missing (E' and E'' are false: $E' = E'' = 0$), rational (noninteger) if E' is true and E'' is false ($E' = 1$, $E'' = 0$; $\frac{1}{4}$ is added), and irrational (in fact, transcendental) if E'' is true ($E'' = 1$, it does not matter whether $E' = 0$ or $E' = 1$; either one adds $1/\pi$ or one adds $1/\pi + \frac{1}{4}$). In this example, it is therefore sufficient to know the outcomes of E' and E'' in order to establish whether X is integer, rational and noninteger, or transcendental; on the other hand, one would need to know whether Goldbach's conjecture was true in order to know whether $0 \leqslant X \leqslant 1$; or whether there were an infinite number of exceptions in order to know whether X were negative ($-1 \leqslant X < 0$); or a finite number (N) in order to know whether $X \geqslant 1$ (more precisely, $N \leqslant X < N + 1$).

Having put forward these examples and discussed them, we can now confine ourselves to cases where X has a physical meaning (or something akin to a physical meaning). To be more specific, we shall deal with those X for which knowledge of their values can only be attained or approached through operations of measurement (which may or may not be precise). A few illustrative examples will make this intuitive explanation clearer; it, in turn, will prove useful for a careful scrutiny of the same ideas insofar as they constitute the standard formulation. We are interested in delving deeply into certain aspects of the latter, but we are even more concerned to warn against the customary readiness to accept that these assumptions can, or must, be assumed valid for 'all' random quantities, with no discrimination. The examples and discussion given above were intended precisely for this purpose.

In order to provide a technical discussion of this topic, we must first of all say what we mean by an 'operation of measurement'. By this we mean an operation which, if applied to X, can only have the effect of restricting the field Q of possible values to an interval[24] (and we repeat that this is an 'experimental' question, not a mathematical one).

These general characterizations only serve as a starting point. It will prove much more important for our analysis to examine closely a few of the main variants from among the many possible. As far as *precision* is concerned, it is of interest to single out the three cases of *bounded, unbounded* and *perfectible* precision; and, moreover, cases of precision which are *possibly perfect* and *certainly perfect*. So far as *partitions* are concerned, we shall distinguish the three cases of *fixed*, *free* and *variable* partitions. It would be pointless and rather tiresome to examine all these variants and their innumerable subcases in detail, not to mention all the other possibilities. What is worthwhile is to exemplify some of them, in order to single out the points which really matter (noting in passing that in this way we obtain an adequate view, even if not complete).

We shall begin with a typical example of a measurement operation with bounded precision and a variable partition. We obtain a value which we know to differ from X by not

23 Which asserts that every even integer can be written as the sum of two primes. Investigation has (so far) revealed no exceptions, but no proof has yet been found, so that there could well be exceptions (even an infinite number of them).

24 It is convenient to fix one's ideas on this case (that of a *two-sided* restriction); it is better, however, not to exclude the limit cases in which the interval becomes a half-line (*one-sided* restriction), or the whole line (*no* restriction; operation failed); or, instead, going to the opposite extreme, into a point (exact measurement). It goes without saying that if $\mathcal{O}$ is not initially the whole line then the above reduce to the intersection of $\mathcal{O}$ with the interval in question (and the operation may have no effect if the actual restrictions obtained are weaker than those already holding for $\mathcal{O}$).

more than a given maximum error, δ. (One could consider, asymmetrically, different limits, $-\delta'$ and $+\delta''$, for negative and positive errors, or one could think of δ – or even of δ' and δ'' – as functions of x, where $X = x$; one could complicate matters even further, by thinking of other random quantities, and so on; however, nothing of conceptual importance is lost, and much is gained in clarity, if we limit ourselves to the simplest case.) Having obtained, by measurement, a certain value x, it tells us that X certainly belongs to the interval $(x - \delta, x + \delta)$. The precision is *bounded* in the sense that we know X up to this δ, whatever it may be. If, however, we have at our disposal operations of this type with δs arbitrarily small, then the operation which consists in first selecting one of them, with a δ corresponding to the degree of precision required, and then in performing a measurement on the basis of it, is an operation with *unbounded* precision (there is always some preassigned margin of error, δ, but this can be chosen arbitrarily small before the measurement is performed).

Even with a fixed δ, we could perform measurements of *possibly unbounded* precision (in the sense that the margin of error might remain equal to δ, but, by chance, could also become arbitrarily smaller). Bearing in mind that all 'assumptions' (like the ones we are about to make) always have an experimental interpretation – even if this is not explicitly mentioned – it is sufficient to assume that it is possible to 'repeat' the previous operation; or, to put it in a better way, to perform another operation with identical characteristics (and we are assuming that there are no constraints of a physical nature, or of cost, administrative veto etc.). If this is done, and two distinct values $x_1 \neq x_2$ are obtained (to fix ideas, suppose $x_1 < x_2$), the two bounds reduce to the single interval

$$x_2 - \delta < X < x_1 + \delta,$$

so that X belongs to the interval

$$x \pm \delta^*, \quad \text{where } x = \tfrac{1}{2}(x_1 + x_2) \text{ and } \delta^* = \delta - \tfrac{1}{2}(x_2 - x_1).$$

Observe that the same conclusion (with more and more possibility of obtaining a small value of δ^*) holds in the case in which one can repeat the same operation over and over again, as many times as one wishes. It is enough to take x_1 to be the minimum and x_2 the maximum of the measurements obtained. Put into words: if one takes the mean of the resulting maximum and minimum, one has a margin of error equal to that of a single observation (δ) less half the difference between the maximum and minimum. Assuming that one could go on to make an *infinite* number of repetitions, the precision is even *possibly perfect* (it suffices to take x_1 and x_2 as the infimum and supremum rather than the minimum and maximum) because it is possible that the difference $x_2 - x_1$ tends to 2δ and, hence, δ^* tends to zero.[25]

25 This could also happen with only a finite number of repetitions, even with just two, if one assumes that the intervals $x \pm \delta$ are to be taken as *closed* (i.e. that the error can reach its upper bound δ). The question is, in itself, rather hair-splitting (given that to have fixed a *precise* δ is already an arbitrary schematization), but it is precisely for this reason that it seems innocuous to follow here the criterion of 'mathematical convenience' (which, in general, we do not approve of). From this viewpoint, it appears preferable to take the intervals to be *open* (and that is why we chose to write them, without comment, in the form $x - \delta < X < x + \delta$, with '<' and not with '$\leqslant$'), thus avoiding the fact that the conventional assumption of a precise separation leads (even though exceptionally) to the possibility of an exact measurement. If the intervals are open, every intersection of a finite number gives an open interval, and hence not a point (neither can it be empty, since we cannot have $x_2 - x_1 > 2\delta$).

If, finally, one proceeds to 'repetitions' of the operation with different precisions, the conclusion does not change in the finite case, nor does it (formally) in the infinite case. The bound can always be written

$$x' - \delta' < X < x'' + \delta'',$$

with $x'' + \delta''$ equal to the minimum (or infimum) of the $x_i + \delta_i$ (and vice versa for $x' - \delta'$). We note, however, that, because the δ_i are also variable, the minimum (and maximum) will no longer be attained, in general, for the i which give, respectively, the minimum and maximum x_i. The interval $x \pm \delta^*$ will be centred at

$$x = \frac{1}{2}\left(x' + x''\right) + \tfrac{1}{2}\left(\delta'' - \delta'\right) \quad \text{with} \qquad \delta^* = \tfrac{1}{2}\left(\delta' + \delta''\right) - \tfrac{1}{2}\left(x'' - x'\right).$$

It is instructive to note that, in any case, δ^* is at most equal to the minimum (or infimum) of the δ_i, given that the interval $x \pm \delta^*$ is contained (as their intersection) in all the intervals $x_i \pm \delta_i$. If, therefore, we have at our disposal operations with δ arbitrarily small, and we can not only choose one (as in one of the cases mentioned above) but can also perform a succession of them with $\delta \to 0$ (or in some way with infimum zero), we have a case of *perfectible* precision if one thinks in terms of performing a finite number of operations, however large (or if one takes into account that at any time point not infinitely far away the number will be finite anyway). We have a case of *certainly perfect* precision if we place ourselves at the end of time, or if we imagine that an infinite number of operations can be performed in a finite time.

We have mentioned precision in its various forms but as yet we have said nothing about the partition. We have merely noted that, in the example considered (and extended in various ways), we had a *variable* partition, since the intervals which could be obtained as bounds could have any end-points whatsoever (with no imposed constraints inherent in the nature of the operations considered). The opposite case – that of a *fixed* partition – is perhaps best illustrated in the case of a number X for which (as in the example of $X = \pi$) one can determine the decimal places one at a time (or take readings from a series of increasingly finely graduated scales). A *free* partition is a partition that one is free to choose from among several others (or, in general, from among an infinite number), but which, after the choice is made, becomes fixed. Think of a measuring scale whose origin and measurement unit can be chosen arbitrarily (for example, instead of X, consider measuring or calculating the successive decimal places of $\sqrt{2} + \pi X$; or measuring angles having rotated the dial of the instrument; or, more generally, to deal in this way with nonlinear functions – for instance, $\log X$, or the tangent of the angle – and so on). A further example occurs with operations for which one can establish in advance the interval to which the question will refer (is it or is it not true that X is in the interval (x', x'')?). If one assumes (certainly) perfect precision, there is nothing else to be said. In general, however, this would not be valid (and, in any case, this is not the form of greatest interest). One can think, for example, that (because it is outside the scope of the apparatus) there may be some doubt as to whether X is near (inside or outside) the end-points, and, in this case, one can imagine that the answer is either YES or NO (by chance), or MAYBE. Supposing that δ is the margin of error (for different δs the same qualifications as were made previously continue to hold) we have the following: in the case of the two answers YES or NO only,

YES means that $x' - \delta < X < x'' + \delta$ and
NO means that $(X < x' + \delta) \vee (X > x'' - \delta)$ (in particular, it is *void* if $x'' - x' < 2\delta$).

In the case of three answers, YES, NO, MAYBE, we have

YES: $x' + \delta < X < x'' - \delta$,
NO: $(X < x' - \delta) \vee (X > x'' + \delta)$,
MAYBE: $(x' - \delta \leqslant X \leqslant x' + \delta) \vee (x'' - \delta \leqslant X \leqslant x'' + \delta)$.

It is perhaps more interesting to consider the simplest case, that of comparison with a single value x, asking whether $X < x$ (the usefulness of this particular case is obvious if one considers that an interval corresponds to two questions of this kind with $x = x'$ and $x = x''$). In this case, the answers have the following interpretations: with the two answers YES or NO,

YES: $(X < x + \delta)$
NO: $(X > x - \delta)$.

(In other words, whatever the answer is, there is no way of excluding the possibility that X can assume any of the values in $x \pm \delta$). On the other hand, with three answers, the MAYBE clearly picks out this possibility;

YES: $X < x - \delta$ (and, *a fortiori*, we certainly have $X < x$),
NO: $X > x + \delta$ (and, *a fortiori*, we certainly have $X > x$),
MAYBE: $x - \delta \leqslant X \leqslant x + \delta$.

In both cases, however, it could happen that the answer MAYBE does not give any information. This would happen, for example, if it could arise both as a result of X being near the end-points of the interval under consideration, and as a result of a chance malfunction of the apparatus. We make this comment not for the pleasure of adding yet another variant, but in order to stress that every question relates to factual circumstances. The logical structures can only be specified with respect to the given circumstances, and not on the basis of some convention fixed beforehand in the belief that one can do without taking these circumstances into account.[26]

Let us now consider the conclusions we can draw as a result of the above discussion. If X is a random quantity, knowledge of which can be attained by means of procedures of the kind considered above, for which sets I can one say that the event $X \in I$ is verifiable? In other words, when is it meaningful, practically speaking, to ask whether X belongs to I?

We have throughout, when introducing a notion, tried to give a realistic analysis of the underlying assumptions (bearing in mind, however, that it would be unrealistic to take these realistic considerations too seriously, thus turning them into metaphysical obsessions). Following this procedure, it would appear in this case that we are justified

26 My insistence on these points (which I hope the reader will excuse) is a consequence of my impression that in general they are not taken sufficiently into account (and sometimes not at all).

in confining ourselves to consideration of just three types of answer. We can label them by again using the terms *bounded*, *unbounded* or *perfect precision* (avoiding the introduction of other terminology, although, for the reasons already given, the trichotomy ignores many further subtle subdivisions).

If we confine attention, for the time being, to the case of a random quantity (geometrically, a straight line – the real line), the three cases can be described in the following way, beginning with the last one.

Perfect precision. This is the case in which one leaves out of consideration the practical difficulties examined in Section 7 and considers the value of X, and hence the question of its belonging to some set I, as perfectly determinable. This is very much a theoretical view but there is no reason to rule it out on these grounds. Indeed, there would seem to be no justification for closing one's mind against any formulation (even those which themselves would lead one to close one's mind against others), unless, in this case, one were taking account of the unrealistic character of such perfect precision, and, for this reason, were to choose to substitute one of the following cases in its place.

Unbounded precision. This is the case in which, by translating into a necessarily idealized form the practical situation that can be realized in the most favourable circumstances, one imagines that, apart from the endpoints, an unambiguous answer can always be obtained to the question of whether or not X belongs to an arbitrary set I. In other words: the question 'is it true that X belongs to the set I?' is partially decidable, in the sense that one can obtain either the answer YES or the answer NO; certainly YES if X is in I, certainly NO if it is not, but it could be either answer, YES or NO, if X belongs to the boundary, $\mathscr{F}(I)$, of I (that is, if every neighbourhood, however small, having X as an interior point, contains both points of I and of its complement $\tilde{I}$). Put another way, having obtained a *measurement* $\hat{x}$ of X, it will be legitimate to conclude that $X \in I$ if $\hat{x}$ is an interior point of I (and so on).

Bounded precision. This is a weakened form of the previous case, obtained by substituting in place of the boundary, $\mathscr{F}(I)$ of I, 'a boundary strip of width δ', which we denote by $\mathscr{F}_\delta(I)$, consisting of the points x for which the neighbourhood $x \pm \delta$ contains both points of I and of its complement. If we call δ-internal (δ-external) those points which belong to I ($\tilde{I}$, respectively) but not to $\mathscr{F}_\delta(I)$, the difference between this and the previous case reduces to using the prefix 'δ-' (we could perhaps even use the term 'δ-boundary'). One could even include the previous case in the present one by introducing a small modification and allowing $\delta = 0$: it would suffice to redefine $\mathscr{F}_\delta(I)$ by replacing 'the neighbourhood $x \pm \delta$' by 'every neighbourhood $x + \delta'$, with $\delta' > \delta$'.

In all cases, this must be interpreted as a possible final result (and not as a possible result of 'an operation' of measurement, which could be improved on by combination with others).

Bounded precision: more general forms, fixed and optional. It is not difficult to weaken this scheme in two different senses, thus obtaining a much closer correspondence to realistic requirements (albeit idealized ones).

Instead of neighbourhoods $x \pm \delta$, with δ fixed, we can define directly $\delta(x)$ as a neighbourhood of x associated with x in an arbitrary way (in general asymmetric, and of

variable length), except for the necessary restriction of reciprocity; $x' \in \delta(x'') \leftrightarrow x'' \in \delta(x')$. For example, we could have neighbourhoods defined by $f^{-1}(f(x) \pm \delta)$, with f increasing, δ constant. This frees us from any particular scale.

A more substantial and indispensable freedom, however, is that we can consider the possibility of obtaining a measurement with precision defined by different laws, $\delta(x)$, a free choice being allowed among them (although – but for the time being we shall not worry about this point – the cost factor may differ). It might be that we are faced with a choice between measuring, with an error $\leqslant\delta$, either X or e^X, or one of three quantities, …, or an infinite number, or X with an error δ (small, but not necessarily arbitrarily small). In such cases, we shall speak of measurements with *optional bounded precision*.

For each $\delta(x)$, the definition of $\mathscr{T}_\delta(I)$ becomes 'the union of all the $\delta(x)$ with $x \in \mathscr{T}(I)$'.

Expressed in this way, it becomes clear how, formally, the notions of the preceding schemes could be directly transported from the one-dimensional case (only a single X under consideration) to the general case (provided that we have a topology). It is, however, necessary to carry out a critical examination – not in the abstract, but from the point of view of the basic issues involved – of whether, and how, essential considerations about observability, similar to those presented above for a single X, can justify the introduction of a topology in the general case. (We did, after all, strive hard to eliminate any trace of topologies which might have arisen naturally, but perhaps, we suspected, irrelevantly.)

8 Continuation: The Higher (or Infinite) Dimensional Case

In the finite-dimensional case the extension of the considerations previously encountered does not involve too many new features. But when do we have a finite-dimensional situation?

Having reshuffled all the points in order to eliminate the topology, the S_r formed by the r random quantities $X_1 \ldots X_r$ *is* merely an infinite collection of points with the cardinality of the continuum, which can be individuated by means of a single number X. (It is not even necessary to have recourse to the Peano curve in order to obtain continuity, since we leave this out of consideration.) We are certainly not saying anything new if we point out that in eliminating the topology we take away the meaning from dimensionality, too; this is a necessary observation, however, if we are to frame the problem as it presents itself in our case.

Let us now make a point in the opposite direction, in order to show that S_r (assuming that we have adopted it in order to have a continuous representation with respect to $X_1 \ldots X_r$) might not be sufficient. If we were also interested in considering the values of another random quantity X, a function of the others, $X = f(X_1, X_2, \ldots, X_r)$, we do not need – logically speaking – a new dimension in order to represent it. If, however, the function f is very irregular – for example, everywhere discontinuous – knowledge of the X_h in the sense of 'unbounded precision' is not sufficient to determine the value of X. In the case of bounded precision, the same difficulty would arise even if X were continuous but varying sufficiently rapidly (for example, $X = \sum_h \sin \lambda X_h$, with $1/\lambda$ small in relation

to the imprecision in the measurement of the X_h[27]). In these cases, if one is also interested in X (and if there is some way of measuring it which is more direct and dependable than calculating it by means of the formula – if not, there is no real interest, only a vain desire), it will be necessary to introduce another dimension for X, in this way going from S_r to S_{r+1} (and so on, in the same way, should we wish to consider several such functions).

Here is another observation, again of a practical nature, which, under different assumptions, reduces the number of distinct dimensions to be considered. Suppose that in the initial formulation one considers a large number of quantities X_1, X_2,..., X_r, like, for instance, the coordinates and velocity components of N molecules (in which case, $r = 6N$). Suppose, also, that, from a practical point of view, we are interested in, and can measure, not the individual X_h (although this would be needed for the theoretical part of the development and calculations) but only some of the macroscopic quantities which derive from them (or, at any rate, some number of quantities r', much smaller than r). In order to study this aspect of the problem, one must refer not to S_r (having the number of dimensions required by the theoretical formulation), but rather to $S_{r'}$ (the space of the r' quantities, X'_k, $k = 1, 2,..., r'$, functions of the original X_h, $h = 1, 2,..., r$, which in practice are either unobservable or irrelevant: $X'_k = f_k(X_1,,,X_2,,,...,,,X_r)$).

It appears, then, that the number of dimensions is also a notion that is neither absolute on the one hand, nor arbitrary on the other. It boils down to being the number of quantities that one requires to measure independently in order that one may know all the quantities of interest to the degree of precision judged satisfactory. Naturally, 'satisfactory' cannot have the meaning of 'to the desired extent' if such a degree of precision is not attainable. The term is then to be understood in the sense of 'we can be content with this, given that other methods of measurement (for example, direct ones) would not increase the precision to the extent required for it to be worthwhile (in terms of efficiency and cost) applying them'. All this depends, of course, on the kind of precision that is attainable.

This 'definition' of the number of dimensions (or, to put it in a better way, this 'suggested way of choosing it') is close in spirit to the necessary extension to this case of topological and allied notions. If by 'a small neighbourhood of a point' (permit me to use the expression) one means, in practice, a 'set of points indistinguishable from it', 'sufficiently close to it compared with the degree of imprecision of the measurement', this indistinguishability must be attributed to the measurements to be made in practice according to the above-mentioned requirements (even if this is a bit extreme compared with the rigid theoretical scheme of some previous formalistic approach).

This having been said in general, there is nothing to add in the case of unbounded precision (any definition of the neighbourhood of a point will do: for example $|X_h - x_h| < \delta (h = 1, 2,..., r)$, or $\sum (X_h - x_h)^2 < \delta^2$). In the case of bounded precision, however,

---

27 The following provides a suitable, simple example. Suppose we are dealing with a phenomenon whose behaviour as a function of time t is given by $f(t) = A(t) + B(t) \sin \lambda t$, where $\sin \lambda t$ represents a daily or annual variation (with period equal to one day or one year), whereas $A(t)$ and $B(t)$ have slower variations (only perceptible on a much longer timescale) because they represent the trend of the mean value of $f(t)$ and the amplitude of its oscillation. In order to investigate its behaviour, it is certainly more instructive to consider and to represent $f(t)$ as a function of two variables, t and $\tau = t - 2\pi k/\lambda$ (where k is the largest integer for which $2\pi k/\lambda \leqslant t$): we then have $f(t, \tau) = A(t) + B(t) \sin \lambda \tau$. (And one day of time τ on the ordinate might appear larger than a year or a century of time t on the abscissa.)

it is necessary to say something more; especially in the *optional* case, which can give rise to a greater variety of circumstances.

In the case of fixed, bounded precision we still have a law δ which to each point $Q = (x_1, x_2,\ldots, x_r)$[28] associates a given neighbourhood $\delta(Q)$ (with the condition that $Q' \in \delta(Q'') \leftrightarrow Q'' \in \delta(Q')$). We shall say that the measurements on the different X_h are 'independent' if the neighbourhood $\delta(Q)$ is the Cartesian product of the neighbourhoods $\delta(x_h)$. If the measurements are not independent, or are performed on combinations of the X_h, or in some other way, the form of $\delta(Q)$ may be anything at all.

This fact, which if δ is fixed has no practical importance, becomes important in the optional case where one can choose, from among various laws, $\delta_1, \delta_2,\ldots, \delta_s$, the δ one prefers (it may be a choice from an infinite number of laws, $\delta \in \Delta$). This possibility of choice opens up the way to much more varied and important features in the higher-dimensional case and it would be pointless to attempt to give a typical example. It will be enough to present a case that throws light on a situation of particular interest, one to which we shall return later from another viewpoint. If X_1 and X_2 are 'complementary' quantities (in a quantum theoretic sense) there exists a relation of a probabilistic nature between the precisions with which one can measure them simultaneously. Translating this (in order to include it within the framework of the present considerations) into a relation which gives bounds, we could say that the margins of error, δ_1 and δ_2, of the measurements of X_1 and X_2 can be chosen at will subject to the restriction $\delta_1\delta_2 >$ constant. As a neighbourhood $\delta(Q)$ (in the (x_1, x_2)-plane) we have the option of any rectangle of constant area (i.e. with vertices lying on a rectangular hyperbola). So far as the question of the number of dimensions is concerned, this circumstance reveals possible causes of uncertainty and complications over and above those already mentioned. In fact, if we consider as adequate the knowledge of X_1 and X_2 with the precision attained by measuring them jointly, we are assuming that we can measure two quantities. If, on the other hand, we are not content with this precision, we can obtain an acceptable measurement for only one of the two quantities. We could choose which one, but it would be only one. Should we, therefore, eliminate the other from the set of quantities to which there corresponds a dimension? The question is (in all probability) just a rhetorical one, because if we do not eliminate the dimension in question it should not cause any extra trouble. Moreover, it is capable of either being excluded or being chosen, and hence it might be called 'potentially observable'. In any case, even these points which we leave open show (to paraphrase Shakespeare) that many more problems arise when we worry about what is 'realistic' than arise if we accept some pre-packaged scheme.

If we wish to go on to the *infinite-dimensional* case, we must again ask ourselves what it means. Again, there does not appear to be a unique answer. We might interpret it in a *weak* sense (thinking of being interested in representing an infinite number of quantities, but only being able to measure an arbitrarily large finite number of them, or combinations of them); or in a *strong* sense (thinking in terms of being in a position to say something which depends on the infinite number of quantities taken as a whole: e.g. to refer to the lim sup etc.).

---

28 We are using r here to denote the 'chosen' number of dimensions; it may be the theoretical r, or any other.

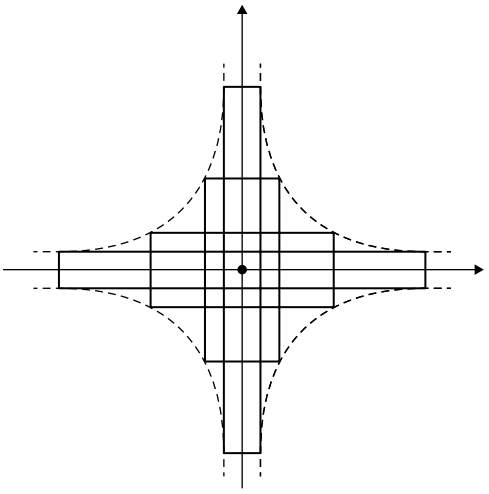


Figure A.1 Rectangles of equal area (concentric, and similarly placed); equilateral hyperbolae, the locus of the vertices.

This latter possibility seems rather theoretical, especially given the discussion of Sections 5 and 6: recall, for example, the discussion concerning the maximum of a function (this was judged to be indeterminable if an infinite number of measurements of all the values assumed were required and determinable if a direct method existed).

The weak interpretation seems to correspond better to the requirements of a theoretical scheme in line with the interpretations conceivable in real terms in practical applications. We shall, as a general rule, stick to this interpretation. For questions where the links with empirical considerations are so tenuous, it is difficult to base one's judgements and decisions on precise reasoning, rather than on impressions. Perhaps the choice here has been dictated by the idea that the weakest formulation is always the most valid one … until such time as it is shown not to be (and, in this case, one will always be able to patch it up after due consideration). The remedy would be more difficult in the opposite case.

Let us recall what is meant by a weak topology in an infinite-dimensional linear space. It is sufficient to require that among the neighbourhoods of a point are the half-spaces containing it. It follows, in fact, that the intersections of a finite number of them are also neighbourhoods (and this applies, *a fortiori*, to sets containing these intersections), and this completes the enumeration of the neighbourhoods. In terms of convergence, this is equivalent to defining the (weak) convergence of a sequence of points Q_n to a point Q by the condition that every linear coordinate of Q_n tends to the corresponding coordinate of Q: $x(Q_n) \to x(Q)$.

In the case of bounded precision, the neighbourhoods will themselves be of the above-mentioned form and we observe, in particular, that they are necessarily unbounded. In fact, they have cylindrical structure: if $\delta(Q)$ is the intersection of n half-spaces (as we can always assume, by virtue of what we noted above) and s is an arbitrary line contained in the intersection of the n hyperplanes which delimit the half-spaces (this intersection is always an infinite-dimensional space; one might say

"all except n"), then together with any point P contained in $\delta(Q)$ are contained all lines through P parallel to s. To visualize this fact more easily, it suffices to note that two planes in S_3 cannot cut all the straight lines (more precisely, they do not cut those parallel to their line of intersection), and the same happens when one considers three hyperplanes in S_4, four in S_5, ..., r in S_{r+1}: *a fortiori*, this happens for r hyperplanes in S_{r+2}, S_{r+3} and so on (and, in fact, the lines which are not cut will determine an infinite number of different directions; ∞^1, ∞^2 etc.). Finally, we note that this happens in any case where a space is cut by hyperplanes which number less than the dimension of the space: all the more so if the hyperplanes are finite in number and the dimension of the space is infinite.

9 Verifiability and 'Indeterminism'

The notion of *indeterminism* and the related notion of *complementarity* have arisen in the context of well-known 'anomalies' encountered in the study of physical phenomena. More specifically, in the study of phenomena where for certain aspects the particle interpretation is appropriate and for others the wave interpretation holds. Neither interpretation can lay claim to being universally acceptable, nor can the two be considered simultaneously without leading to 'contradictions'.

The question has been, and continues to be, a live topic of discussion; many-sided, and requiring special competence in several fields. The arguments put forward have offshoots in many directions, making it extremely difficult both to encompass them all (even if one restricts oneself to the essential points) and to single out with sufficient clarity either one topic, or a small group of them, on which one would like to concentrate attention.

The aspect which concerns us here is the logical-probabilistic one (and, in fact, for the time being, just the logical aspect, although with a view to the probabilistic side of things, for which it will serve as support). The study of this aspect could not be carried out, however, without touching upon points relating to other aspects and without indicating the position taken up with respect to them, a position which appears to correspond to that underlying the proposed choice of approach in the logical field.

Let us, without further ado, indicate which works we shall be referring to most frequently in what follows: on the one hand, that of von Neumann, and, in particular, the exposition and development given by Bodiou, whose formulation is in the field of direct interest to us; on the other hand, that of Reichenbach, who seems to me to present the questions most lucidly from the logical and philosophical point of view.[29] The solution

29 John von Neumann, *Mathematical Foundations of Quantum Mechanics*, Princeton University Press (1955) (a translation, by Robert T. Beyer, of *Mathematische Grundlagen der Quantenmechanik*, Springer, Berlin (1932)); J. von Neumann and G. Birkhoff, 'The logic of quantum mechanics' *Annals of Math.* (1937); G. Bodiou, *Théorie dialectique des probabilités* (etc.), Gauthiers-Villars, Paris (1964); Hans Reichenbach, *Philosophical Foundations of Quantum Mechanics*, University of California Press (1944).

References to these works in the present section will be indicated by means of initial and page number; R, p. 238, for example, for Reichenbach.

that I will put forward is a different one, but it could, in a certain sense, be seen as a simplified version of those set out by these authors.[30]

The different solutions, or interpretations, relating to the points we have to consider here, are concerned, explicitly, with the systematization in the logical domain of those kinds of statements which, because of their association with 'anomalies' like those mentioned above, lead to confusion. In order to incorporate them, it is suggested that, in general, we must have recourse to new logical structures, different from the usual structures, such as many-valued logics, or logics with modified operations and rules (in particular, 'nonmodular' logics).

The very starting points on which the analysis of these problems is based differ, however, one from the other. This difference is mainly between those who consider the problems as strictly peculiar to quantum physics, and who therefore pose the problems directly in terms of its technicalities, and those who see the problems as problems of thought in general. In the latter case, these problems could still appear more or less bound up with quantum physics, but only for contingent reasons; that is because they satisfy needs which actually arise in that theory (some would say 'exclusively' so, some 'mainly').

The formulation of von Neumann (vN, pp. 247–254) is strictly in terms of quantum theory, and takes as its starting point a Hilbert space (of functions ψ) in which the (linear Hermitian) operators correspond to quantities. An *event* is a quantity capable of assuming only the two values 0 and 1,[31] and therefore represented by a *projection-operator* E (which is idempotent, $E^2 = E$; that is having possible values – and eigenvalues – either 0 or 1); that is by a closed linear manifold $\mathscr{M}$ (that onto which E projects orthogonally). The event E is certain or impossible according to whether ψ belongs to $\mathscr{M}$ or is orthogonal to it; in all other cases, E has probability equal to the square of the projection of ψ onto $\mathscr{M}$. Two events are incompatible if they are orthogonal; they are simultaneously verifiable (not 'complementary') if they are commutative (in which case the logical product and logical sum are meaningful); and so on. To quote von Neumann (p. 253):

> 'As can be seen, the relation between the properties of a physical system on the one hand, and the projections on the other, makes possible a sort of logical calculus with these.'

---

30 My attitude had previously consisted in rejecting ad hoc interpretations in relation to quantum physics in order to reduce everything, essentially, to familiar situations (to facts which were 'complementary' in the sense that they were conditional on mutually exclusive experiments; like the behaviour of an object in two different destructive testing situations; or the victory of a tennis player in two different tournaments taking place at the same time in two different countries). A mention of this can perhaps only be found in the CIME (Centro Italiano Matematico Estivo) course given in Varenna, 1959. This solution seems to coincide with that of B.O. Koopman, *Quantum Theory and the Foundations of Probability* (1957).

Subsequent reflection (after a good deal of reading – the most relevant being that mentioned above), has not changed my original view, but rather made it more precise. In any case, it is, of course, simply an attempt at explanation (as we remarked at the beginning of the chapter) given the many issues involved, some of which may have escaped my notice.

31 Let me just mention, as an interesting curiosity, that this is the same convention as I had adopted (in a paper of 1964, and now here) after much hesitation, considering it novel and perhaps unacceptable. I subsequently realized that, far from being new, it had been in use since 1932 (together with all its developments). I wonder if the fact of its not being taken up confirms my doubts about its unacceptability?

The study of this kind of logical calculus (in terms of projections) has led (vN and B) to the identification of *nonmodularity*[32] as the characteristic property which distinguishes the lattice of this logic from that of standard (Boolean) logic.

A development which is inspired by the trend towards studying, in a more autonomous manner, or even completely separately, logic (and probability) on such a lattice – or on similar structures, also referred to many-valued logics – can be found in the work of Bodiou. His intentions are clearly summarized in the following passage (B. p.7):

> 'The primary motivation for our work, quantum theory, might appear contingent and particular, and capable of disappearing by the wayside if, by chance, quantum theory should come to be incorporated within a classical theory, which eliminates its "anomalies". This is what contemporary probabilists seem to believe and to expect. We shall attempt to show that they are wrong, and that the quantum calculus is simply a special case, imposed by necessity, of a general calculus of probability, which we call *dialectic*. This latter, far from being an unnatural growth on the body of the classical calculus, in fact subsumes it.'

Discussions of which statements and interpretations 'are or are not meaningful' are more directly considered, and more rigorously set out in Reichenbach, in a form which makes specific reference to quantum mechanics (and compares, in this context, the work of various authors), but which, from a conceptual point of view, can be adapted to any context whatsoever. For this reason, we shall develop our own analysis by using his (Reichenbach's) remarks as a guideline, putting forward our remarks as comments on his. In any case, the object of the analysis is that of finding the logical constructions that will prove suitable for resolving the difficulties in which we find ourselves; a topic which has attracted many currents of ideas from many different sources. This goal does not appear to have been achieved, nor does it appear that the efforts to reach it have opened up any promising avenues. I have the feeling that (as I said in my preliminary remarks in Section 6) the correct path is straightforward and simple, but it is my belief that it is obscured precisely by preconceived ideas about what it is that constitutes a necessary prerequisite for any logic.[33]

This would also appear to be a move in the direction of a natural continuation of a natural process; that of eliminating the drama from the initial state of confusion brought about by the appearance of something new, in contrast to one's accustomed way of seeing things. This has already happened for the Copernican system and for non-Euclidean geometry, for logical paradoxes and for relativity theory, and was bound to happen for the 'anomalies' of quantum physics. It seemed as if either *logic* itself was on trial or had

---

32 See, for example, L. Lombardo-Radice, *Istituzioni di algebra astratta*, Feltrinelli, Milan (1965), pp. 332 ff.

33 A comment seems called for at this point. My agreement or disagreement with the opinions of various authors concerns the individual points which necessarily arise in the course of an argument, and does not indicate any general position for or against. In every work there are inevitably a number of points with which the reader agrees or disagrees, either strongly or to some extent, or is in doubt about, or indifferent to, or simply does not understand. This also holds true for those works which I value to the extent of making them the basis for a discussion, an indication in itself of the stimulation I derived from them.

fallen apart completely. Reichenbach makes it clear, however, that logic, including probabilistic logic, is not to blame; on p. 102 he says:

> 'The rules of logic cannot be affected by physical experiences. If we express this idea in a less pretentious form, it means: If a contradiction arises in physical relations, we shall never consider it as due to formal logic, but as originating from wrong physical interpretations.'

The attributing of 'anomalies' not to the *structure of the physical world*, but rather to the 'structure of the languages in which this world can be discussed' is even more decisive. 'Such analysis expresses the structure of the world indirectly but in a more precise way' (R, p. 177). These are the languages which, by means of *definitions*, introduce into the world of observable phenomena something that we might call 'interphenomena' (non-observables). As examples of such 'definitions', consider those which attach to the 'observed value' the meaning of being the value of the quantity *before and after*, or only *after*, and so on. As examples of such languages, consider the particle language, the wave language and a neutral language: The first two '… show a deficiency so far as they include statements of causal anomalies, which … can be transformed away, for every physical problem, by choosing the suitable one of the two languages. The neutral language is neither a corpuscle language nor a wave language, and thus does not include statements expressing causal anomalies. The deficiency reappears here, however, through the fact that the neutral language is three-valued; statements about interphenomena obtain the truth-value *indeterminate*' (R, p. 177). The same situation is described by Bodiou as the existence of several 'coherent formal systems', like the mechanics of points, and the wave theory of light. Two 'attributes' pertaining to different systems, like a statement in particle form and one in wave form, 'might be *incoherent* without being *contradictory*' (B, p. 11).

The appearance of the word 'indeterminate', or of the distinction between 'incoherent' and 'contradictory', indicates that in order to find something suitable for our purpose we must bring into the world a new logic. There is no difference, in principle, between the approaches of the two authors cited. The one introduces straightaway a third 'logical value' and then goes on to define the logical operations by means of 'truth tables'; the other defines the operations axiomatically and could (it seems to me – in fact, he does not, although I think he should do so) define the 'truth values' on the basis of them.

Reichenbach distinguishes two variants depending on whether a statement that is neither *true* nor *false* is called *indeterminate* or *meaningless*. The different names do not correspond to different meanings of the partitions into the three cases; the change in name corresponds to the case in which 'it is necessary to make an observation H in order to know whether E is true or false'. One agrees then to say that

E is *true* if observation H has given the result E,
E is *false* if observation H has given the result not-E
E is $\left\{\begin{array}{l} \textit{meaningless} \\ \text{or } \textit{indeterminate} \end{array}\right\}$ if the observation H has not been made.

Using the notation introduced for conditional events, this turns out to be exactly what we agreed to say by writing $E|H$ instead of E, and putting $E|H = 1$ (true), or $= 0$ (false), or $= \varnothing$ (void) (and, if we wish, we could call it 'meaningless' or 'indeterminate' instead of 'void').

The meaning of the trichotomy does not depend at all on which words we use; the way in which it is defined is the only thing that matters. It might be conceded, however, that it does make some difference whether we use 'meaningless' instead of 'indeterminate'. It is a difference of philosophical attitude – an acceptance of the Bohr or Heisenberg interpretation. And there are formal consequences if one believes that a meaningless statement cannot even be mentioned, whereas by calling it *indeterminate*, and considering this, as '*an intermediate truth-value*' (R, p. 145) lying between true and false, it becomes permissible to speak of it and *above all to work with it*.

This is, in fact, the requirement that must be satisfied if something is to be called a mathematical structure or, in particular, a logical structure. It is very easy to construct such a structure. Considering the tableau on the left, there are 3^9 (=19 683) different ways of substituting the letters T, F, I (True, False, Indeterminate) in place of the asterisks and one can choose a subset of these to which to assign the title of operation and a symbol (e.g. +) to replace the symbol ° between A and B at the corner.

$A \circ B$	B: T	F	I
A: T	*	*	*
F	*	*	*
I	*	*	*

$A + B$	B: T	F	I'
A: T	T	T	T
F	T	F	I
I	T	I	I

The table headed $A + B$ is thus filled with entries that are either T or F or I, placed in the 1st, 2nd, or 3rd row according to whether A is true, false, or indeterminate, and in the 1st, 2nd or 3rd column depending on the value of B (an example is given in the tableau on the right). Reichenbach (p. 151) introduces seven such binary operations (some of which are taken from the work of Post): disjunction and conjunction (extensions of logical sums and products), three forms of implication (the standard one, an alternative and a form of quasi-implication), two forms of equivalence (the standard one and an alternative), and three unary operations of negation (cyclical, diametrical and complete). Four of these operations are due to Reichenbach himself, but he leaves out 'some further implications' defined by Post. Variants due to other authors are also mentioned.

Without going into a more detailed discussion (which would lead on to more substantial objections), we note that all this could be expressed in terms of two-valued logic by thinking of a 'three-valued event' (in our terminology a 'conditional event', but nothing is altered if one refers to it – or thinks of it? – in a different way), E, say,

expressed in the form $E'|E''$, or in that of the partition into three cases in which it is either true, false or void:

$$E^T = (E = 1) = E'E'', \quad E^F = (E = 0) = \tilde{E}'E'', \quad E^I = (E = \varnothing) = \tilde{E}''.$$

An E whose logical value depends on the logical values of two other 'three-valued' events, A and B, is obtained by defining as logical functions of A^T, A^F, A^I, B^T, B^F, B^I, both E' and E'', and E^T, E^F, E^I (should they turn out to be exhaustive and exclusive), and putting $E = E'|E''$ or $E = E^T|(E^T + E^F)$.[34] In this way, one avoids creating a number of symbols and names of operations and consequent rules (which are difficult to remember and sort out, and difficult to use without confusion arising). Above all, one avoids creating the tiresome and misleading impression that one is dealing with mysterious concepts which transcend ordinary logic.

Bodiou (like, of course, many previous authors whose approach and notation he follows) does not base his work on three truth-values (although, in his notation, the three possibilities for a proposition a seem to me to be expressible in the form a, Ca, $C(a \vee Ca)$, where C denotes negation). He does not even have a symbol for the 'third value' (whereas he uses u, True, and $\varnothing$, False, corresponding to our use of 1 and 0). This takes one even further from an immediate understanding of the meaning. There is also a section (B, pp. 30 ff) devoted to many-valued logics in which there are M 'truth-values', which can be denoted by $k/(M - 1)$, $k = 0, 1, 2,\ldots, M - 1$, but not even here is there a value equivalent to 'Indeterminate'. The work, in fact, proceeds in an entirely different direction, in which the value $W(a)$ of a proposition should have a meaning something 'similar' to 'probability' (but with its own 'rules': $W(a \vee b) = W(a) \vee W(b)$, with the same for $\bigwedge$; fortunately, we have $W(Ca) = 1 - W(a)$).[35]

With this, the time has now come (for two reasons which are formally identical, but as far as interpretation goes are totally unrelated) to examine the real meaning of these questions; no longer just formally, but in depth. And it is at this point that, in our examination, we must take into consideration, together with the notion of indeterminacy, the notion of *complementarity*.

10 Verifiability and 'Complementarity'

The essential problem, the basic doubt which came to the surface in the previous analysis, can be expressed formally in the following way. Suppose we have two or more 'three-valued events' (we shall consider them as conditional events, but it does not make any difference), and let us denote them by $E_1 = E_1' \mid E_1''$, $E_2 = E_2' \mid E_2''$, …, $E_s = E_s' \mid E_s''$. Can it be meaningful and interesting to define another such, an $E = E'|E''$, whose meaning is related to the meaning of the others? And, in this case, will its 'truth-value' be a function of those of the others?

If we think of the general case (for example, of s conditional bets) it is likely that a few events (simple, two-valued ones) will be of interest; like $E_1^T E_2^T \ldots E_s^T, E_1^T + E_2^T \ldots + E_s^T$,

34 Note that E'' is the same thing as $E^T + E^F$, whereas for E' it does not matter whether we take E^T or $E^T + E^I$, or any event in between ($E' = E^T + D$, with D contained in E^I).

35 Such rules, proposed by other authors, are changed by Bodiou in a way which brings them closer to probability theory (but then, of course, *one no longer has operations on logical values*).

and the similar forms with E_h^F or E_h^I (expressed by means of simple events), which express the fact that either all the events, or at least one, are won, or are lost, or are called-off (because the conditioning event did not occur).

If we think of cases of actual interest, however, it appears that we have to reconsider our whole approach to the problem, since some Es may arise that are connected with the E_h in a meaningful way, although not necessarily with their logical values. As a trivial example – but, for this very reason, instructive – let us begin by considering the (two-valued) events E_h', and for each of them construct the (two-valued) event E_h'' = 'I know (at this moment) whether E_h' is true or false'. By this means, we have transformed our field of events into a field of three-valued events, in which the third value stands for 'unknown', whereas True and False stand for 'known as true' and 'known as false'. One often emphasizes – and for good reason – that 'Indeterminate' is not to be confused with 'unknown' (see, e.g., R, p. 142). The same thing can be underlined in a rather better way by saying that the two notions coincide only in this particular example. One might say that this example corresponds to thinking of what 'I know at this moment' as frozen (I will no longer be able to learn anything about that which is now unknown to me, and I have no further interest in it; for me it will be for ever indeterminate, or even 'meaningless').

In this example it is natural to call the logical sum of two conditional events E_1 and E_2, $E = E_1 \vee E_2$, the conditional event corresponding to the logical sum $E_1' \vee E_2'$ of the corresponding events E_1' and E_2'. We have, therefore, $E = E'|E'' = (E_1' \vee E_2')|E''$, where E'' = 'I know (at this moment) whether $E_1' \vee E_2'$ is true or false' (i.e. if at least one of E_1' and E_2' is true), and this does not coincide with $E_1'' \vee E_2''$ (although it necessarily contains it) since one might well know, for instance, that someone 'arrived yesterday or today', but not know which. It follows, therefore, that the event $E = E_1 \vee E_2$ thus defined is *not* a logical function of E_1 and E_2 in the sense we have seen so far (function of their logical values). E is certainly true if at least one of the E_h is (certainly) true, certainly false if they are both known to be false; but if they are both indeterminate (unknown) it could be either indeterminate or (known to be) true. (And if there are more than two of the E_h, one has the latter case if at least two are indeterminate and the others false.)

This example, as we have said, is trivial; but the cases in which considerations of this kind find an actual important application are precisely (I would even say exclusively) those modelled on the same scheme (except that 'I know …' is replaced by 'it has been verified that …', or 'it will be verified that …' – within a certain time period, for example – and so on).

Quantum theory provides an obvious example of a case in which everything is more clear-cut. If E_1 and E_2 denote two events (or, equivalently, the respective projection-operators), an observation by means of the operator $E_1 \vee E_2 = E_1 + E_2 - E_1E_2$ is an observation for the event-sum, but not for either of them individually. Similarly, one can make an observation of $X + Y$ or XY and so on without making observations of the two separate random quantities X and Y. The concept is an analogous one but, if we wish to confine ourselves to projection-operators representing events, we must restrict ourselves to considering the events consisting of whether or not X (or Y, or $X + Y$, or Y etc.) belongs to a given interval, $E = (a \leqslant X \leqslant b)$; that is $E = 1$ if X lies between a and b, and $E = 0$ otherwise.

When we turn to verifiability (in the various senses considered in previous sections) the situation is similar. Leaving aside the details and the finer points (we do not have to

repeat them here, having dwelt upon them – perhaps at too great a length – already), it will suffice to refer to the 'trivial example' considered above, taking as 'indeterminate' that which will be 'known (not now, but) after a certain time, or after certain checks have been made, or after having obtained some given information, and so on.' Here, too, '*E* is indeterminate' is the statement of an objective fact; the lack of the information required in order to decide the truth or falsity of *E* within the specified time period and according to the rules laid down for doing so. The difference is that the assignment of the value 'indeterminate' (and the same for 'true' and 'false') is, in this sense, not immediate (as it was in the trivial example): it is not excluded, however, as in two-valued logic (where 'unknown' is always considered as a temporary state of affairs, pending knowledge of the truth – even if the period of waiting should turn out to be in vain, or to last for ever).

Within this context, it appears to be possible to pose the problem of complementarity and to discuss its real meaning. Let us first of all observe that, in the sense we have just used it, the qualification 'indeterminate' might be attributable to an event *as of now*: that is when, on the basis of what we already know about the events, and about the possibilities and known means of obtaining information, we are in a position to exclude the possibility of getting to know whether *E* is true or false within the given time period, and according to the specified mode of doing so. Clearly, only a few of the 'unknown' events (or *possible* events, as we used to call them) will be 'as-of-now indeterminate' (in exceptional cases all of them might be and this would reduce to the trivial example considered above). We could say, in order to be more precise, that the division (at a given moment) of the events into certain, impossible and possible (i.e. with values known, as of now, to be true, false or unknown) could be pursued further, subdividing the possible (unknown) events into five subcases depending on the situation considered as 'final'. Specifically, we can distinguish the events which will eventually be certainly indeterminate (the case *I* already mentioned), or those for which there is doubt between the outcomes *T–I* (true–indeterminate), *I–F* (indeterminate–false), *T–F* (true–false), *T–I–F* (true–indeterminate–false).

From these conclusions concerning single events, we can pass to the properties of two or more possible events. When we were restricting ourselves to True and False, for example, we were able to say whether two or more events were incompatible or exhaustive. This meant that although it was possible for any of the events to occur, not more than one of them actually could. In the same way, it is possible that in the case of indeterminacy similar exclusions can be made. It may be certain, as of now, that, from among two or more events, each of which might or might not in the end turn out to be indeterminate (i.e. they are all either *T–I*, *I–F* or *T–I–F*; none of them are *I* or *T–F*), at least one remains indeterminate, or at least one does not, and so on, and so forth.

The interesting case in practice is that of two events, one at least of which remains certainly indeterminate (but it is not known which; otherwise it is easy to couple it with another one). Two such events are called *complementary* (and a similar definition holds for quantities, as we shall shortly see). The purpose of the more general discussion given above was merely to show how the notion corresponds to a natural examination of the possibilities that present themselves when we extend the classification by introducing 'indeterminate' as a third logical value.

Complementary events arise, for example, when establishing:

whether or not a tennis player wins if he takes part in one or other of two tournaments taking place at the same time in two different countries;
whether a coin will show Heads or Tails the next time it is tossed, assuming that the next toss is performed by either Peter or John;
what are, fixing them in one's mind, the registration number and the features of the driver of a suspect car that flashes past (assuming that it is at best possible to observe one or other of the two items);
what is the behaviour of one and the same object when it is subjected to one or other of two destructive tests;
whether a given building (e.g. the Tower of Pisa) will remain standing until some specified date under the assumption that some kind of repair work is carried out (or assuming some other project); and so on.

As well as events, one could equally well speak of *complementary quantities*. Referring to the above examples, we could consider the remaining life of the Tower of Pisa conditional on one or other of the hypotheses considered (of which only one will be observable – that conditional on the course of action actually chosen). Two random quantities X and Y are, by definition, noncomplementary (i.e. simultaneously measurable) in the strict sense, if this condition holds for all the events $X \leqslant x$, $Y \leqslant y$, for arbitrary x, y (and, in quantum theory, this reduces to the spectral decompositions of the corresponding operators; see vN, Chapter II, and, for noncomplementarity, p. 254). An amusing example, but one which well conveys the idea (in a nutshell), is the complementarity of the two measurements that a tailor would have to make simultaneously when one of them requires the client to hold his arm straight downwards, and the other requires that he hold it parallel to the floor and with the elbow bent to give a right angle.

The most celebrated example is undoubtedly that of complementarity in quantum mechanics, and there is no doubt that this is the most important case, because of the profound nature of the implications regarding our conception of the nature of phenomena and the knowledge of them that we can attain. There is also a more 'technical' and precise way of expressing the condition of complementarity for events in this case. As we have already mentioned, E_1 and E_2 are noncomplementary events if, as projection-operators, they commute. Does this (together with related factors) provide sufficient justification for the idea that one has to make a *distinction of a logical nature* between complementarity in the realm of everyday affairs and in that of quantum physics? The answer would seem to be no. Otherwise, why should we not say that incompatibility – corresponding as it does in the quantum theory formulation to orthogonality – should be considered in that context as something completely different, even though it is exactly the same thing? This comment is certainly not sufficient to settle the argument; nor is the much more basic fact that we have up to now presented the notion of complementarity without having encountered any need to introduce such distinctions. We shall have to be more specific, albeit in a summary fashion, about the physical meaning of the problems under consideration, while, on the other hand, we must examine, from a critical standpoint, the arguments put forward on the basis of these physical considerations to support the opposite point of view.

As a first step, we shall simply develop the *description* of what we shall need as background for our purposes. This can then be used as a basis for the understanding of the logical situation and also by those who are not familiar with the physical and mathematical interpretations of the schemes on which we shall base our considerations. We shall take as our starting point the remarks of Section 9, concerning the interpretation of events as projection-operators and, in fact, we shall restate this, for the convenience of the reader, and provide an integrated and extended version by including the case of quantities. Notwithstanding the inevitable fact that many things will be glossed over, this should be sufficient, and the resulting picture should turn out to be clear and precise enough for the purposes for which it is intended.

11 Some Notions Required for a Study of the Quantum Theory Case

As the fundamental notion, we take a space $\mathscr{H}$ which, by means of its points, or rather vectors, provides a suitable representation of the 'states' in which a given physical system S can find itself. The space is like ordinary three-dimensional space, with all the affine and metric properties (those of analytic geometry) but is infinite-dimensional (*Hilbert* space). Its points, or vectors, represent functions (functions defined on the space of possible configurations of the system; for example, of three coordinates x, y, z in the case of a single free particle, and of $3N$ in the case of N distinct particles). The 'state' of the system, at a given instant, is characterized by one of these functions, its ψ-function (or ψ-point, or ψ-vector, as we call the point, or vector, which represents it in the space $\mathscr{H}$); ψ is such that the vector has modulus $\|\psi\| = 1$.

The way the system evolves in time is described by the variation of ψ as a function of time (deterministically set out by equations similar to those of classical mechanics). The difference is that these equations no longer tell one how the configuration of S varies, as it is and as it is observed, but only how the probability of finding it in this or that configuration varies if one submits it to an 'observation' at some future time t.[36]

---

36 Let us just mention some of the omitted details. The functions ψ (and, in general, all the functions considered) are complex (roughly speaking, for the same reason as it is convenient to express oscillations in terms of $e^{i\omega t}$ rather than $\cos \omega t$), and, as such, considered as vectors, they have bounded moduli ($f^2 = \int |f|^2 \, dS < \infty$). The space of these functions is the *Hilbert space* with the *Hermite* inner-product ($f \times g = \int f.g^* \, dS$, where the asterisk denotes the complex conjugate; we always have $(f \times g) = (g \times f)^*$ and $|f \times g| \leqslant ||f|| \; ||g||$). One can directly characterize an infinite-dimensional linear metric space as a *Hilbert* space by adding the properties of completeness and separability. With a system of (orthogonal, etc.) Cartesian coordinates, it is the space of points defined by sequences of coordinates $x_h (h = 1, 2,\ldots, n,\ldots)$ such that $\sum_h |x|^2 < \infty$ (and this expression gives the modulus of the vector with coordinates x_h; the Hermitian inner-product of two vectors is given by $\sum_h x_h y_h$). The linear operators that we shall come across are also Hermitian (or self-adjoint: $A^* = A$, where A^*, the adjoint of A, is defined by $A^* f \times g = f \times Ag$); the operators can be represented by matrices (with reference to an orthogonal Cartesian system) with entries A_{rs} (and then $(A^*)_{rs} = (A_{sr})^*$; A is Hermitian if $A_{sr} = (A_{rs})^*$). See vN, especially pp. 34–46, and for a more direct exposition and interpretation, E. Persico, *I fondamentali della meccanica atomica*, Zanichelli, Bologna (1936).

Let the above serve to give an idea of the various detailed specifications which, like the present one, would be out of place in the main text if they were to give the reader the impression that he has to acquire a knowledge of these notions, or to refresh his rusty memory, or to worry about the details, in order to understand those few points on which his attention would be better focused. And let it serve also as a warning for those who were tempted to accept the present formulations literally, or to be put off at finding them incomplete.

We cannot 'see' the vector ψ; it is a mathematical abstraction in a space that is a mental fiction. But, by starting with the results of the last observations made (and assuming that the system S has not, in the meantime, been subjected to any external disturbance), and knowing the laws governing its evolution, we can, in principle, determine it. In any case, assuming that we knew the vector ψ, little or nothing would be known about the actual configuration of the physical system S in which we are interested, and about its evolution.

More precisely – in order to make clear what ψ is or is not sufficient to explain – let us consider an arbitrary event E, that is any statement whatsoever concerning the space S at a given instant (this must always be understood, even if not mentioned explicitly; two events or two quantities of the same kind, but relative to different times, are two distinct events or quantities). What we can say about E is that it is certainly *true* if the vector ψ belongs to some given linear manifold $\mathscr{M}$ (associated with E), and certainly *false* if it belongs to the linear space of all vectors orthogonal to M. In these two cases it is superfluous to make an observation, because the answer would certainly be the one we have just given (but it would also be innocuous because it would not disturb the system at all). If, on the other hand, the vector ψ is neither contained in nor orthogonal to the space $\mathscr{M}$ an observation concerning E gives an unforeseeable result. Knowledge of the state, which is all contained[37] in the vector ψ, does not determine this result in advance, but it is not without value, because it gives all that can be given: that is, the *probability*. The details are as follows: we decompose the vector ψ into two components, one parallel to $\mathscr{M}$, the other orthogonal; that is into $\psi = E\psi + \tilde{E}\psi$ (in this way indicating the projection E onto $\mathscr{M}$, and the orthogonal projection, $\tilde{E} = 1 - E$, onto H – $\mathscr{M}$). The probabilities of the E and $\tilde{E}$ that result are given by the squares of the respective projections:

$$\mathbf{P}(E) = E\psi^2, \quad \mathbf{P}(\tilde{E}) = \tilde{E}\psi^2.$$

Note that, instead of $||E\psi||^2 = E\psi \times E\psi$ (1st form) we can also write $E\psi \times \psi$ (2nd form, which equals $E\psi \times (E\psi + \tilde{E}\psi)$, whereas $E\psi \times \tilde{E}\psi = 0$), which is also valid even if E is not a projection-operator and has a similar meaning even in this case (as we shall see).[38] In the case where E is a projection-operator, one verifies immediately that $\mathbf{P}(E) + \mathbf{P}(\tilde{E}) = 1$, as was necessary. In fact (because of the orthogonality of the two components, and hence by Pythagoras), we have

$$\mathbf{P}(E) + \mathbf{P}(\tilde{E}) = E\psi^2 + \tilde{E}\psi^2 = \psi^2 = 1.$$

The interpretation of E as a projection-operator turns out to be even more expressive, however, in light of the following. The most important fact is that the vector ψ is not restricted to a passive rôle of indicating the probability of the required result being observed. The observation itself is forced to choose whether to fall into the space $\mathscr{M}$

37 See the comments that are made later following the discussion of the possibility of explanations introducing 'hidden parameters'.

38 The equivalence between the two forms does not hold in the general case. There, in fact (see footnote 36), we have $A\psi \times A\psi = A^*A\psi \times \psi$, and, in the case which interests us (the Hermitian form), this equals $A\psi \times \psi$. In order that it equals $A\psi \times \psi$, we must have A idempotent; i.e. A = projection-operator (with all the eigenvalues idempotent, $\lambda = \lambda$; that is, $\lambda = 1$ or $\lambda = 0$).

corresponding to E, or into the orthogonal space $\mathscr{H} - \mathscr{M}$ corresponding to $\tilde{E}$ (and it does so with the probabilities indicated). The outcome simply constitutes the information about the choice made. From the position taken up after the jump we have obliged it to take, the system returns to an evolution according to the previous laws until there is a new disturbance.

This picture (anthropomorphic, but this perhaps helps one grasp the ideas in the absence of a more detailed technical exposition) contains 'in a nutshell' all that is required for a complete treatment. It will suffice, essentially, to consider simultaneous questions about several events, rather than a single one (this will also hold for measurements of quantities), and to distinguish the cases which give rise to the circumstances to be discussed.

Instead of a partition into two opposite events (E and $\tilde{E}$) we can think of a 'finer' partition, still 'complete', into some number n of incompatible events, $E_1, E_2,\ldots, E_n$, or even into an infinite number. Each E_h will be defined by the corresponding (closed) linear space $\mathscr{M}_h$, and all these spaces must be taken to be orthogonal to each other, and such that taken altogether they form the entire space $\mathscr{H}$ (i.e. there must not exist a vector in H orthogonal to all the $\mathscr{M}_h$, and then there will not exist vectors which are not linearly dependent on the vectors of the $\mathscr{M}_h$). The total dimension is countable, being coincident with that of $\mathscr{H}$. It follows that, in the case of a finite partition, at least one of the $\mathscr{M}_h$ must be infinite-dimensional (in particular, note that for E and $\tilde{E}$ either one is infinite-dimensional or both are). In the case of a countable partition, this is not necessarily the case, and we can even consider the extreme case where all the spaces $\mathscr{M}_h$ are one-dimensional.

This is the fundamental case in terms of which the discussion of all the others is framed. In other words, it is the case in which we have a system of orthogonal Cartesian axes corresponding to an infinite set of events $E_h (h = 1, 2,\ldots, n,\ldots)$, interpretable, for example, as the (distinct) values, λ_h, which a quantity Z can assume: $E_h = (Z = \lambda_h)$. On the other hand, we clearly have $Z = \sum_h \lambda_h E_h$ (as a random quantity), since one and only one of the E_h will occur (and will take the value 1; all the others will be 0), and the sum will reduce to the corresponding value λ_h. Defining Z as an operator (associated with the quantity of the same name) in the same way, it seems clear from the identity of the written forms that one is dealing with the operator formed by multiplying the axis vectors (functions) E_h (the eigenvectors or eigenfunctions) by the λ_h (the eigenvalues). This gives the prevision (or mathematical expectation) of the quantity Z by means of the same formula used for E (2nd form):

$$Z\psi \times \psi = \left(\sum_h \lambda_h E_h\right)\psi \times \psi = \sum_h \lambda_h \left(E_h \psi \times \psi\right) = \sum_h \lambda_h \mathbf{P}\left(E_h\right) = \mathbf{P}\left(Z\right).$$

In a similar fashion, one obtains, immediately, the distribution function of Z: putting $E_z(\lambda) = \Sigma E_h(\lambda_h \leqslant \lambda)$, one has the event $E_z(\lambda) = (Z \leqslant \lambda)$, or the related projection-operator; it follows that

$$E_z\left(\lambda\right)\psi \times \psi = \mathbf{P}\left(E_z\left(\lambda\right)\right) = \mathbf{P}\left(Z \leqslant \lambda\right) = F_z\left(\lambda\right),$$

where we denote by F_z the distribution function of Z.

The collection of projection-operators $E_z(\lambda)$ (or the set of their linear spaces, each of which, if we proceed in the direction of increasing λ, contains all the preceding ones)

defines the spectrum of Z; in this case, a discrete spectrum (there are a countable number of values of λ_h).

Going back to the physical problem, we can repeat what was said in the case of an event. If ψ belongs to one of the axes, and only then, the corresponding event E_h will certainly be true; that is, thinking of the quantity Z, its value will be λ_h with certainty. It is unnecessary (but harmless) to make an observation. In all other cases, the vector ψ will be forced (if we make an observation on all the E_h together, i.e. on Z) to choose along which axis it is to lie: the result (E_h, or λ_h) indicates which choice was made, and, after the jump, we go back to the normal evolution.

There is one difference, and a very important one, with respect to the previous case. Now we know exactly,[39] after the observation, the position chosen by ψ (whereas, beforehand, we knew only that it belonged to $\mathscr{M}$, or alternatively to $\mathscr{H} - \mathscr{M}$). This happens only in the case now under consideration; that of a partition which gives rise to spaces which are all one-dimensional. When one defines on them a quantity Z, it is necessary that the λ_h values (the eigenvalues) are all distinct (simple), otherwise, the refinement of the subdivision we have reached will in part be destroyed (one has the case of 'degeneracy'). For the case we are dealing with (the nondegenerate case) one says – for obvious reasons – that a 'maximal observation' has been made.

We now turn to the problem of complementarity of observations.

Can we ask for the simultaneous verification – that is with one and the same observation – of two or more events? Or for the measurements of two or more quantities? (And we note that 'simultaneously' can only mean 'with one and the same observation'; another observation made immediately afterwards would already find ψ changed by the effect of the first one.)

The answer is obvious if one thinks of two 'maximal observations', like the measurements of two quantities Z' and Z'', to be performed simultaneously. In doing so, we force ψ to lie on one of the axes of the first system and also on one of the second system. Now ψ obeys any order whatsoever, but cannot accept contradictory orders – and this would be the case if the two systems of axes do not coincide. In such a case, Z' and Z'' cannot be measured simultaneously; that is they are *complementary*.[40] If the axes do coincide, the result is trivial, because Z' and Z'' are functions one of the other (if Z' assumes the value λ'_h, it means that the hth-axis has been chosen, and hence that Z'' takes on the value λ''_h). The coincidence of the system of axes implies commutativity (in terms of operators, the condition is $Z'Z'' = Z''Z'$), and the same also holds in the case of events, $E'E'' = E''E'$, or of non-maximal quantities, $XY = YZ$. Non-maximal quantities X, Y relating to one and the same system of axes (suppose it to be that of Z) are obtained by taking eigenvalues μ_h and ν_h, which are not all distinct (so that X and Y as functions of Z, $X = f(Z)$, $Y = g(Z)$, are not invertible). Conversely, if X and Y are not complementary, and have as possible values the μ_i and ν_i, respectively, one can construct a Z (corresponding to a maximal observation) of which X and Y are functions, and having distinct values λ_h corresponding to all the compatible pairs (μ_i, ν_j). In particular, in the case of events, noncomplementarity, $E'E'' = E''E'$, means that the corresponding spaces $\mathscr{M}'$ and $\mathscr{M}''$ are mutually orthogonal (i.e. if we call M the intersection, $\mathscr{M} = \mathscr{M}' \mathscr{M}''$, then two vectors,

39 The ψ is, in fact, uniquely determined, because a multiplicative constant (real or complex) is irrelevant.

40 Think of the example of the tailor: complementary measurements are those that to be made simultaneously would require the client to simultaneously assume several different, incompatible positions!

one from $\mathscr{M}' - \mathscr{M}$, and one from $\mathscr{M}'' - \mathscr{M}$ – i.e. from $\mathscr{M}'$ and $\mathscr{M}''$, respectively – and orthogonal to M, are always orthogonal to each other; then, in fact, and only then, is the product of the two projections the projection onto the intersection M, and does not depend on the order). As special cases, one has the case of inclusion (if $\mathscr{M}' - \mathrm{M} = \{0\}$, we have $\mathscr{M}' \subset \mathscr{M}''$, $E' \subset E''$), and that of incompatibility ($\mathscr{M} = \{0\}$, $E'E'' = 0$ = impossible).

Noncomplementarity between X and Y can be interpreted in the same way because it is equivalent to the noncomplementarity of each of the events (projection-operators) $E_x(\mu)$ and $E_y(\nu)$; in other words,

$$E_x(\mu)E_y(\nu) - E_y(\nu)E_x(\mu) = 0$$

for any μ and ν whatsoever. Geometrically, this is equivalent to the orthogonality of the spaces $\mathscr{M}'_\mu$ and $\mathscr{M}''_\nu$ (i.e. orthogonality between the vectors of $\mathscr{M}'_\mu - \mathscr{M}'_\mu\mathscr{M}''_\mu$ and of $\mathscr{M}''_\nu - \mathscr{M}'_\mu\mathscr{M}''_\nu$). One could consider the same condition in a weaker version (limiting ourselves to checking the validity for certain values of μ and ν instead of for all of them), but we shall consider this in the context of 'continuous spectra', where it is more interesting.

The case of the 'continuous spectrum' arises with a quantity that can assume any value (between $-\infty$ and $+\infty$, or in an interval etc.), rather than just a finite or countable set of values as considered so far. In quantum physics one deals with nonquantized quantities (like the coordinates) in addition to the quantized ones (like energy).

In this case, too, in considering quantities X, Y,..., everything can be expressed by $\mathrm{E}_x(\mu)$, $\mathrm{E}_y(\nu)$,..., except that an $\mathrm{E}_x(\mu)$ will actually vary for all increments in μ (and not just in going through certain values of μ; the eigenvalues $\mu = \mu_i$) and the distribution function

$$F_x(\mu) = \mathbf{P}(X \leqslant \mu) = E_x(\mu)\psi \times \psi$$

will, in general, turn out to be continuous.[41] A decomposition into a finite or infinite number of incompatible events could be obtained by dividing the axes of the μ in some fashion into intervals $\mu_i < \mu < \mu_{i+1}$ ($i = 0, \pm1, \pm2,\ldots$, in order to denote them in increasing order, and letting them be, in general, unbounded in both directions). Only in this way can we have a partition

$$E_i = (\mu_i < X \leqslant \mu_{i+1})$$

that gives us, in the way we indicated previously, an 'approximate measurement' $\hat{X}$ of X, defined by choosing a subdivision μ_i, and in each interval a value $\hat{x}_i$, and then setting $\hat{X} = \sum_i \hat{x}_i E_i$. In other words; $\hat{X}$ is the function of X defined by the step-function

$$f(X) = \sum_i \hat{x}_i (\mu_i < X \leqslant \mu_{i+1}).$$

From this point of view, X is not a measurement with absolute precision, but with arbitrarily high precision if we substitute for it an $\hat{X}$ defined by a function with arbitrarily small steps. An observation on $\hat{X}$ forces ψ to lie in one of the spaces $\mathscr{M}_i$ (corresponding to $E_i = E_x(\mu_{i+1}) - E_x(\mu_i)$) and it is never maximal because the subdivision could always be made finer.

---

41 We ignore the mixed case of probability in part concentrated, in part continuous, etc. (see Chapter 6, 6.2.2–6.2.3).

If we now consider X and Y (both with continuous spectra), the condition for noncomplementarity is still the commutativity of X and Y as operators, $XY - YX = 0$; in other words, commutativity between the $E_x(\mu)$ and the $E_y(\nu)$. If the condition holds, then, apart from obvious complications, what was said in the case of the discrete spectrum also applies here: for example, it is also true in this case that one can construct a Z of which X and Y are functions, but that one can only obtain it (clearly) by means of procedures based on the Peano curve or something similar (vN, p. 178).

If, however, we content ourselves with approximate measurements, $\hat{X}$ and $\hat{Y}$, then the orthogonality of $E_x(\mu_i)$ and $E_y(\nu_j)$, relative to the points of subdivision chosen for the μ and ν, is sufficient for noncomplementarity; this weaker condition may also hold if X and Y are complementary. In other words, complementarity does not necessarily exclude the possibility of simultaneous approximate measurements; that is of two suitably chosen $\hat{X}$ and $\hat{Y}$.

And at this point we arrive at the special case of quantum mechanics, where complementarity often arises in the particular guise of noncommutativity, expressed by

$$XY - YX = h / 2\pi i \quad \left(h = \text{Planck's constant}\right).$$

This holds where X and Y are coordinates, and for a conjugate impulse, or, more generally, in the terminology of classical mechanics, for 'canonically conjugate' quantities.

From this relation of noncommutativity (and hence complementarity between X and Y) we can derive a justification of Heisenberg's Uncertainty Principle, which indicates the way in which the precision of the measurements of X and Y – which can be made arbitrarily high if performed separately – turn out to have a reciprocal relationship under a simultaneous observation.[42]

The following is just a brief development of this crucial point. From $XY - YX = a$, it follows that

$$\left(XY - YX\right)\psi \times \psi = a\psi \times \psi = a\psi^2 = a \ \left(\text{if}\, \psi = 1\right).$$

We note also that

$$XY\psi \times \psi = Y\psi \times X\psi \leqslant Y\psi \cdot X\psi,$$

with a similar result for YX. It follows (by the triangle inequality!) that the bound for the difference is given by

$$\left(XY - YX\right)\psi \times \psi \leqslant 2.\left\|Y\psi\right\| \cdot \left\|X\psi\right\|,$$

that is

$$\left\|Y\psi\right\| \cdot \left\|X\psi\right\| \geqslant \frac{1}{2}|a| = \frac{h}{4\pi}.$$

42 The procedure which is briefly outlined here is taken from vN (p. 230, ff.), and is there (note 131, p. 233) attributed to ideas of Bohr, and work of Kennard and Robertson.

The inequality holds no matter where we take the origin for X and Y, and, in particular, if we take it at the mean value. The two moduli then have an interpretation as standard deviations, and we have the uncertainty principle in its usual form: $\sigma_x\sigma_y \geqslant h/4\pi$.

This means that it is impossible to approximate X and Y by means of some $\hat{X}$ and $\hat{Y}$ by choosing arbitrarily small 'steps' for both of them: their order of magnitude must be such that the product (as an order of magnitude) is not less than h. Geometrically, the subdivision into rectangles in the (X, Y)-plane, a consequence of the subdivision by means of the μ_i and ν_j, respectively, on the x-, y-axes, cannot be so fine as to give rectangles whose areas have orders of magnitude less than h (the choice of the ratio of height to width remaining arbitrary). These rectangles are the regions for which it can be 'verified' whether the pair of measurements fall inside or not: $(\hat{X} = \hat{x}_i)(\hat{Y} = \hat{y}_j)$ is, in fact, equivalent to $(\mu_i < X \leqslant \mu_{i+1})(\nu_j < Y \leqslant \nu_{j+1})$. Observe that this is precisely one of the conditions of 'bounded precision' considered in Section 8 (Figure A.1).

On the basis of the digression, which we are now about to end, we might, at this point, take up again the discussion of the logical aspects of indeterminacy. However, let us first take advantage of the opportunity that has grown out of the discussion concerning the precision of a measurement in the quantum theoretic field, in order to examine the question in relation to the considerations we put forward about the subject in general (in Sections 7 and 8). We have just pointed out the similarity between relations of indeterminacy and bounded precision in terms of area; more fundamental, however, is the analogy between the case of 'unbounded precision' (considered in Section 7) and the situation presented (following von Neumann) for the measurements of nonquantized quantities (with continuous spectra). These quantities (vN, p. 222) '… could be observed only with arbitrarily good (but never absolute) precision', in contrast to what happens with the '… introduction of an eigenfunction which is "improper", that is, which does not belong to Hilbert space', a procedure which '… gives a less good approach to reality than our treatment here. For such a method pretends the existence of such states in which quantities with continuous spectra take on certain values exactly, although this never occurs.' These critical comments are directed towards procedures that make use of the Dirac function (and they are repeated very frequently). In this connection, I think it appropriate to indicate its relation to the point of view that we are following, both to avoid misunderstanding and to make things clearer. I sympathize with von Neumann's attitude, not when he seems to be inspired by scruples of mathematical rigour and attacking imprecise definitions (because, generally speaking, formal imperfections can always be removed), but when he shows care in not attributing absolute certainty and precision to a quantity without really good reason. I approve even more strongly of the observation (as he adds in note 126, p. 222) that not even attributing X to one of the intervals of the subdivision can be considered as certain, except as an idealization: 'Nevertheless,' he concludes, in an admirably undoctrinaire manner, 'our method of description appears to be the most convenient one mathematically at least for the present.' 'I sympathize', also, in the sense that I would like quantum physics to make room for this elegant example of contraposition; of quantities that are or are not quantized, which are, respectively, precisely measurable or not. In any case, it cannot, of course, be a matter of taste, whether mathematical or philosophical, and if the opposite formulation should, on a closer examination, turn out to correspond more closely to a meaningful physical interpretation then it must be welcomed with open arms.

12 The Relationship with 'Three-Valued Logic'

Let us now go back to three-valued logic and to the related conceptual questions that are raised by quantum mechanics.

So far as the nature of 'three-valued logic' is concerned, we came to the conclusion that the 'three values' correspond well to the requirements of applications – quantum-theoretic or otherwise – but that they do not give rise to a 'logical calculus', because the most meaningful considerations are not connected with operations that could be performed on such 'values'. If one examines the actual situations directly, the pre-conceived choice of a machinery consisting of formal operations similar to those of ordinary logic does not appear to be appropriate as the unique way of constructing a formulation which is to replace it.

The best proof is provided by the many-valued logic – 'similar to the calculus of probability' – which Bodiou mentions and develops. In order to have the satisfaction of finding a true relationship (in the calculus of probability), he has to put together two things which are falsely defined (in the calculus of probability) and which, on the other hand, cannot be modified if one wants them to be expressed as 'logical functions' (and the final consolation lies in the observation that they only work if one has either probability 0 or 1, in which case formal, two-valued, logic, without probability, is essentially sufficient).[43]

The most important point to be examined is the reason behind the different attitudes (already mentioned in Section 10) of those who regard 'Indeterminism' as a concept specifically and exclusively belonging to quantum physics, and those who see no distinction *of a logical nature* between this case and the world of everyday affairs (although nobody denies the very important and significant differences which derive from the physical and mathematical structures peculiar to the quantum-theoretic set-up).

Von Neumann, in speaking about the 'sort of logical calculus' to which the projection-operators gives rise, says of this calculus that '… in contrast to the concepts of ordinary logic, this system is extended by the concept of "*simultaneous decidability*" *which is characteristic for quantum mechanics*' (vN, p. 253).

It seems that such sentences have no practical implication and, therefore, no actual content. So far as von Neumann is concerned, it may be that he never examined the possibility of examples of a different kind. This is not the case, however, with Reichenbach (as we shall see); indeed, one might think that he was constantly preoccupied with such

---

43 Perhaps this 'many-valued logic' might be useful in other cases, and in other senses, without reference to probability. For example, by giving a proposition a a certain value $V(a) = k$, if, in a system with a given set of ordered axioms $A1, A2, \ldots, An$, the proposition is decidable (true or false) on the basis of $A1, A2, \ldots, Ak$, but not using only $A1, A2, \ldots, A(k-1)$. The scheme as it stands does not even work in this case, but, in a certain sense, we get closer.

So far as non-modularity is concerned, one can observe that modularity no longer holds in our scheme, $(E|H)$ (or in similar ones), when the 'truth-value' (Void or Indeterminate) is considered greater than 0 (False) and less than 1 (True), and a scale of intermediate values (in various possible senses, e.g. probability), which are considered not comparable with Ø, are inserted between these two values. The most natural convention would be that of taking $\mathbf{P}(E|H)$ as the value, putting $\mathbf{P}(E|H) = Ø$ if $H = Ø$; possibly using 0^* and 1^* to distinguish the certainly False and certainly True cases (obtaining the partially ordered set of values $0^* < 0 \leqslant p \leqslant 1 < 1^*$, $0^* < Ø < 1^*$, Ø not comparable with values $0 \leqslant p \leqslant 1$). It is not clear to me whether this has any connection with the appearance of non-modularity – in many different ways, some not immediate – in the treatments given by von Neumann and Birkhoff, and Bodiou.

dilemmas (like waves versus particles, values before and after etc.). These dilemmas were, instead, clearly resolved by von Neumann, as is shown, for example, by the following remark (note 148, p. 282):

> 'In contrast with this, however, it is to be noted that quantum mechanics derives both "natures" from a single unified theory of the elementary phenomena. The paradox of the earlier quantum theory lay in the circumstance that one had to draw alternately on two contradictory theories (electromagnetic theory of radiation of Maxwell–Hertz, light quantum theory of Einstein) for the explanation of the experience.'

The attitude of Bodiou – see the previous quotation (B, p. 7) – also seems to stem from having overcome these distinctions. On the other hand, it seems an inevitable progression to attain more and more comprehensive views, which remove from their isolation those things which, when they first appeared, seemed abnormal.

What is the difference then, from a logical point of view, between the complementarity or noncomplementarity of measurements in the case of a physicist and in that of the tailor (whom we met above in our trivial example)? Or among the examples given previously – like that of the coin whose next toss could be made by either Peter or John – and an example of a quantum-theoretic nature? It is precisely in the context of such an example that Reichenbach has developed his arguments (R, pp. 145–146, and p. 168), basing himself upon an absolutely rigid division between the indeterminacy of the quantum world and the determinacy of the macroscopic world. This division is so complete that Reichenbach says the following concerning the outcome of *that toss which John might have made* (but which instead was made by Peter). Since it is a question of 'a macroscopic affair, we have in principle other means of testing', by making precise measurements of the state of John's muscles before or after the toss made by Peter, and in many other ways:

> '… or let us better say, since we cannot do it, Laplace's superman could. For us the truth value of John's statement will always remain unknown; but it is not *indeterminate*, since it is possible in principle to determine it, and only lack of technical abilities prevents us from so doing.'

In discussing the merits of the question, one might object that the 'determinism' of the macrocosm – to which Reichenbach makes explicit reference – has a merely static character and this renders completely unpredictable those facts for which numerous microscopic circumstances might prove decisive (not to mention the fact that even the result of a single collision between particles, as recorded on a photographic plate, can cause macroscopic phenomena like the publishing of papers, the holding of lectures and conferences, and endless indirect consequences and repercussions). Moreover, not even Laplace (so far as I know) ever suggested that his 'superman' was capable of predicting not only everything that is going to happen but also what would happen if … something that is not going to happen were to happen. How could it come about that the state of the muscles, and so on, could inform us about the result of the toss that has not been performed (and why not the text of a conversation that has not taken place; the adventures of a journey not undertaken; etc.), rather than informing us directly that it is

predetermined that the toss, or the conversation, or the journey, will not take place (or did not take place)? To my knowledge, no-one, even in theological discussions, has ever claimed to have decided whether divine omniscience includes the knowledge of what exactly would have happened to the world conditional on every imaginable hypothesis about the form of Cleopatra's nose (or any other fact, either substantial or irrelevant, concerning the world's history).

In my opinion, however, there is no point in entering into the merits of such questions, physical or metaphysical as the case may be, because logic can only be *neutral* and *anterior* with respect to any contingent circumstance of scientific knowledge, or hypothesis, concerning the world of phenomena.[44] Logic has to be applied to the wider field of everything that is imaginable, and the inevitable circumstance that fantasy is of so little use in extending the field beyond what has already been observed or realized is already too restrictive. Science fiction itself has rarely anticipated reality by more than a few decades. To make use of new ideas or discoveries can be legitimate for the purpose of bringing up to date points of view in logic by including in its domain new areas of what is conceivable, areas which had previously been ignored (and this is what we are attempting to do). The approach, which consisted, instead, in making every logical theory restrictive and ephemeral, by reducing it, moment by moment, to a reflection of current scientific views, would have got things upside down.

Before turning to another topic, it would perhaps be appropriate to clarify certain views on the theme of determinism, given the connection with discussions pertaining to the present theory, and given that we have commented upon it (even though in order to decide that it was not relevant). In my opinion, the attachment to determinism as an *exigency of thought* is now incomprehensible. Both classical statistical mechanics (or Mendelian hereditary) and quantum physics provide explanations – in the form of coherent theories, accepted by many people – of apparently deterministic phenomena. The mere existence of such explanations should be sufficient to give the lie for evermore to the dogmatism of this point of view. What I mean is that the fact that such theories exist, or are conceivable, should be sufficient (no matter if they are wrong, or even if they are merely successful mental constructs, clearly science-fictional in character).

It is a rather different matter to pose oneself similar questions from what one might call a psychological–aesthetic angle, rather than a dogmatic one. As a result of our own individual tastes and habits, each one of us will have a propensity to find one or other of a deterministic or indeterministic formulation of a law or theory more or less simple and convincing. In particular, we evaluate with greater or lesser sympathy (*a priori* – i.e. before some possibly deeper knowledge or examination of the detailed reasons for and against) the ideas which tend to characterize probabilistic-type quantum theory as merely a partial explanation, unsatisfactory and provisional, and requiring replacing sooner or later by something deterministic.

Personally, I am of the opinion that nothing should ever be excluded *a priori*: tomorrow's notions will almost certainly be as inconceivable for us today as today's notions would have been for a man of the nineteenth century, or for Neanderthal man. This is, however, a distant prospect; the foundations of physics are those we have today (perhaps for many decades, perhaps centuries) and I think it unlikely that they can be interpreted

44 On the other hand, this has been perfectly expressed elsewhere by Reichenbach himself.

(or adapted) in deterministic versions, like those that are apparently yearned for by people who invoke the possible existence of 'hidden parameters', or similar devices. I hold this view not only because von Neumann's arguments against such an idea seem to me convincing (vN, pp. 313–328), but also because I can see no reason to yearn for such a thing, or to value it – apart from an anachronistic and nostalgic prejudice in favour of the scientific fashion of the nineteenth century. If anything, I find it, on the contrary, distasteful; it leaves me somewhat bewildered to have to admit that the evolution of the system (i.e. of its functions ψ) is deterministic in character (instead of, for example, being a random process) so that indeterminism merely creeps in because of the observation, rather than completely dominating the scene. This can actually lead one to search for some meaning which makes the function ψ objective, although this notion is the very least suited to appear to be capable of such a transformation.

In any case, for what concerns us as human beings, interested in foreseeing the future with some degree of confidence on the basis of our scanty, imprecise and uncertain knowledge of the present and the past, all arguments about determinism are purely academic and have no more meaning than would a discussion about the number of angels that can dance on the head of a pin. No matter how the world's history develops, nobody could disprove either the assertion that everything is determined by the past through iron laws (but we can foresee either nothing or very little because we are too ignorant both of the past and of the laws), or the assertion that everything occurs 'by chance' (and this does not exclude the possibility that 'by chance' things might develop according to some 'law'). In the final analysis, it seems to be of very little consequence or assistance to us whether we take up a position for or against the plausibility of the hypothesis that Laplace's superman could work out the entire future *if only he knew the entire present in every detail*. Such a statement, in the sense we have just examined, must, in fact, be said to be neither true nor false, but instead indeterminate, the hypothesis being, without any doubt, illusory, and therefore false.

13 Verifiability and Distorting Factors

As the final part of our survey of the various factors that are important when we attempt to determine and verify the outcome of an event, it remains to consider the most troublesome of them. These are the factors which, for reasons relating to the individuals involved, or to their self-interest, are capable of modifying the outcome, or of influencing its verifiability, or of simply raising doubts about the possibility of such distortions.

Many examples of this are well-known, and we shall just quote them without having anything useful to say about overcoming the difficulties. The deepest discussion (which may well be new) will centre, however, on the events of the three-valued logic illustrated in Sections 10–12 above, where it seems impossible to give a complete definition without encountering similar difficulties regarding possible distortions. Let us begin, however, with the most well-known and obvious cases.

These are events for which the will of the individual concerned enters in directly (and, in this respect, there is nothing to distinguish this case from that of events which depend on animals, or on other natural factors). There is a difference – a distorting factor – when such a will can be influenced by facts which are objects of our study, and which therefore alter this very object of study. This happens in evaluating a probability if the

circumstances upon which the evaluation itself is based are modified by the evaluation, or the knowledge of this evaluation, or a contract drawn up on the basis of this evaluation, and so on.

This much is true: although we are again speaking of probability before the appropriate time, we have to do so in order to present the examples; the object under consideration is, however, the difficulty of avoiding the problems by means of detailed specifications in the description of an event.

The evaluation of the probability of an event can influence its occurrence. If someone, at some given moment, perhaps because of a vague feeling, or even for no reason at all, considers the danger of a traffic accident to be higher than usual, he will try to be more careful, and the risk will diminish. If, on the other hand, we are dealing with an event whose positive outcome is desired – like succeeding in a business deal, an examination, or a race – it can happen that a greater feeling of confidence leads to one being in a better position to succeed.

The knowledge of someone else's evaluation of probability can have a marked snowball effect, as a result of the confidence that tends to be placed in the opinions of experts. If in circles which are considered well-informed the expectations are pessimistic (or optimistic) and an increasing number of people, when informed of these opinions, behave as if they correspond to reality, the expectations will end up by being borne out by reality – even if they were initially without foundation.

Nevertheless, the most direct example is provided by the influence of a person's self-interest on the outcome of events. In the case of insurance, it can lead to faked or fraudulent accidents; but we are still dealing here with the kind of case which is, to some extent, identifiable. Much worse (from a logical point of view, since it constantly eludes one's grasp) is the effect of the insufficient precautions that an individual might take, knowing that he is insured. Similar influences are at work if a prize is attached to the occurrence of an event (e.g. an additional bonus for each goal, either for the player who scores, or for the whole team to share out), or even if it arouses admiration, or merits reproach.

It is even easier for a person to have an influence on the process of verification rather than on the outcome itself. An individual who is interested in proving that an event has occurred will devote a lot of energy to obtaining information to this end, and will take care to collect the necessary documentation and to send it to the appropriate authority. On the other hand, someone interested in concealing such news will be more or less negligent, or may even attempt to suppress it, or to destroy the evidence.

In order to avoid all this, one should provide a description of the event which is sufficiently detailed to preclude the possibility of distortion. In fact, the clauses of an insurance policy abound in detailed specifications of the obligations of the person concerned, the risks excluded, and so on (although it is clearly not possible to extend the specification beyond those cases which are easiest to define and to pick out[45]).

An even more entangled situation is to be found in the theory of games. In the simplest case, one has two players, each of whom must make a decision (without knowing the decision of his opponent), and the result (one player's gain, the other's loss) depends on the two decisions made. Each would like to know the decision of the other in order to

45 A discussion, together with useful examples, can be found in H.M. Sarason, 'Come impostare e applicare le statistiche assicurative', *Giorn. Ist. Ital. Attuari*, I (1965), pp. 1–25.

adjust his own decision accordingly; not knowing it, he could evaluate the probabilities of the various decisions the other might make, and in order to do this he would need to go through a similar reasoning process by putting himself in the other's shoes.

This and other more complicated situations are objects of study in games theory. But all the aspects of distorting factors that we have mentioned so far are only intended as remarks in passing, merely to put the reader on guard against the difficulties one encounters when dealing with cases where they arise (the difficulties may or may not be serious, but they are virtually impossible to eliminate).

Above all, these examples serve as an introduction, in order that it should not appear (misleadingly) to be a rather special feature which arises when one delves more deeply into the study of 'three-valued' events. There is, in fact, a particular, novel feature in this case, but it arises later, and is not related to the distorting factor (which derives from the choices that can affect verifiability). We shall examine this latter aspect first.

A conditional event $E|H$ presents no problem of this kind if E and H turn out to be known with certainty – as true or false – within the time and manner specified. In fact, if we think in terms of having made a bet – our guideline – we will then know, without any room for doubt, that it is called off if H turns out to be false, won if both H and E turn out to be true, lost if H but not E turns out to be true. But what if H or E, or both, turn out to be nonverifiable (in some preassigned manner; for example, within a given time period during which the bet has to be decided)? As a first step, we must decide what is to happen to the hypothetical bet in such circumstances. It seems natural – and, in any case, this is what we shall do – to make the following convention: it is either won or lost, respectively, only in the cases of H and E true, H but not E true, respectively; it is called off both if H is false, and if H is indeterminate, and also if H is true but E is indeterminate. In formal terms, considering E and H as three-valued events, $E = E'|E''$, $H=H'|H''$, the conditional event $E|H = (E'|E'')|(H'|H'')$ would correspond (in Reichenbach's terminology) to *quasi-implication* (as introduced by him), with the following truth table (in Reichenbach's notation, $E|H$ corresponds

$E\|H$		H		
		T	I	F
E	T	T	I	I
	I	I	I	I
	F	F	I	I

to $H \ni E$). In terms of the four simple (two-valued) events E', E'', H', H'', this becomes

$$E\,|\,H = (E'\,|\,E'')\,|\,(H'\,|\,H'') = \begin{cases} 1 \quad (T) & \text{if} \quad E'E''H'H'' \\ \varnothing \quad (I) & \text{if} \quad \sim(E''H'H'') \\ 0 \quad (F) & \text{if} \quad \tilde{E}'E''H'H''; \end{cases}$$

in other words,

$$E\,|\,H = (E' \mid E'')\,|\,(H' \mid H'') = E' \mid (E''H'H'').$$

Distorting factors enter in here, too, as soon as one allows the possibility that somebody might influence the outcome, or the knowledge of the outcome, of $E|H$. The particular case of greatest specific interest is that in which H represents the performing of the experiment – or, more usually, experiments – from which information about the outcome of E is drawn (either in fact, or potentially). This covers the cases of all measurements and experiments in both classical and quantum physics, and of all the investigations that are appropriate for the ascertaining of the truth of any assertion concerning practical matters. This situation arises most clearly when E consists just of the result of an experiment which must be expressly performed (e.g. H = the toss of a coin, or the launch of a satellite, and E = Heads, or entering into orbit, respectively). In such a case, it would make no sense at all to enquire whether E was true or false without assuming that H were true; but the case in which E is thought of as being true or false independently of an experiment H for ascertaining it no longer appears different when one is concerned with the actual ascertainment of E. We can very well imagine that E = A.N. Other has gone down with a certain illness, or E = the residue of a substance contains poison, are statements that are true or false in themselves, independently of the fact of our knowing whether they are true or false. If we, or anyone, wish to know whether E *is* true or *is* false (and not merely to say that it is one or the other) then E should be replaced by $E|H$, where H denotes the performing of an act leading to its ascertainment. We could say, for example, that H = A.N. Other undergoes tests to establish whether or not he is infected with the given disease, or H = the residue of the substance analysed in order to ascertain whether or not it contains poison, and then E = the outcome is positive. But what tests and analyses should be performed?

Let us exclude the possibility that an experiment (e.g. the tests or analyses mentioned in the above examples) could give a wrong answer: this would by no means be absurd, because experiments concerning facts related to the one we wish to ascertain can only lead us to increase or decrease the probability we attribute to it. Everything stems from our convention of considering that a question has been answered only if it is certain, and we call E indeterminate if the ascertainments that have been made have not proved sufficient to resolve the doubt. (Just as, in the case of 'insufficient evidence', it would be inadmissible to claim that a suspect is both guilty and not guilty.)

However, it is only rarely that by performing an experiment H one obtains an answer with certainty. What usually happens (at least in cases which are sufficiently complicated for it to be worthwhile to apply considerations of this kind) is that H may give an answer (in which case it is an exact answer), but, on the other hand, may not (and then E remains indeterminate). To be precise, H should not simply denote the performance of some given experiment, but rather the successful performance of it (in the sense that, with respect to E, the answer is either YES or NO, and not MAYBE). If we wished to split hairs, we could put $H = K'|K$, where K denotes the experiment in the sense of its performance, K' the fact that K was a success, and thus

H is the successful performance of K (i.e. the hypothesis that ensures the ascertainment of the truth or falsity of E).[46]

In general, however, there will be no one unique experiment K which we can (or cannot) perform in order to ascertain E. There will exist various possibilities K_1, K_2, K_3,... (and even if there were only one 'type' of experiment available, we could always vary the time, the apparatus, or the experimenter etc.) and they might or might not be compatible for various reasons (and possibly not repeatable in the case of failure), ranging from physical incompatibility to contingent limitations (e.g. lack of time, apparatus or available personnel, funds, raw materials etc.). In order not to further complicate the notation, we can suppose that the list given by the K_i includes not only the individual experiments (e.g. let K_1, K_2,..., K_{38} denote the performance of just one of 38 possible different experiments), but also all possible combinations or strategies (e.g. the one consisting in first performing K_5, K_{19} and K_{22}, and then, if none of these succeeds, K_7, and, if this is still not sufficient, K_9 and K_{31} together, and then stopping whatever happens, is a strategy which will be denoted by a number greater than 38 – e.g. by K_{728}). For each K_i, K_i' will mean that K_i was successful; that is that K_i succeeded in establishing whether E was true or false. In the case of an individual experiment, K_2, say, K_2' will mean that this experiment was successful; in the general case, for example for K_{728}, K_{728}' will mean that at least one of the component experiments of the strategy was successful (and – in the case of a sequence of experiments, as in the example of K_{728} – the experiments following will then not even be performed). By going into more and more detail (like the time and manner of performing the experiments, possible repetitions etc.), the number of distinct strategies could be increased without limit. So far as our notation is concerned, however, it is simply a question of extending the list of the K_i.

In this way, the problem of ascertaining E is translated, in practice, into the problem of ascertaining one of the conditional events $E|K_i$, depending on the choice of K_i, which is arbitrary within the limitations imposed (of time, money etc.). And it is thus that the arbitrariness has its effect on the verification of E. In extreme cases, there might be one-way experiments; that is experiments which either prove that E is true, or prove nothing (or vice versa). Suppose that one experiment shows whether or not a given liquid is pure water, and another experiment shows whether or not it contains strychnine. If the question is whether or not the liquid is poisonous, the first experiment can only return a negative answer, and the second only a positive one (because to know that it is not pure water, or that it does not contain strychnine, neither proves nor disproves the presence of poison). Even without taking these extreme cases into consideration, any method might present, by its very nature (and taking previous experiences into account, according to the evaluation of each individual), different characteristics in its functioning, and different probabilities of breaking down, depending on whether E is true or false.

Up to this point, we have merely been dealing with cases involving distorting factors, just like the others considered previously (even if these cases deserve special attention because they are less obvious than most other examples). The most important specific

46 Only thus can one avoid having to consider E itself (as well as $E|H$, which becomes, in this notation, $E|K$) as a three-valued event (indeterminate, notwithstanding the – unsuccessful – performance of the experiment). We have our doubts about the actual utility of such notation, other than for a once and for all explanation, and we avoid insisting on, or taking up a position, regarding the desirability of more or less logically perfect forms of notation.

factor in the present case is, however, quite a different one, which – as we mentioned at the beginning – only 'arises later', after the analysis given above of the experiments K_i, their successes K_i', and the consequent realization of the corresponding hypothesis H_i.

The new factor is the following: to be realistic, one should also substitute E_i for E. Let us explain right away what we mean by this. If we perform the experiment K_i, its success K_i' does not give us directly the answer 'E is true', or 'E is false'; it does not make directly visible to us the fact that we wish to affirm or deny by these phrases. Neither, if we are dealing with the more general question of measuring a quantity, does it enable us to realize what the value is by making it visible or tangible. The answer reduces to a signal (a movement, a light, a noise, a colour etc.; in the case of quantities, the position of a pointer on a dial, the reading of a counter, the height of a column of mercury etc.). For an event, we shall have one of two signals, E_i or $\tilde{E}_i$, as possible outcomes of the experiment K_i (as well as the absence of any answer at all – or, if one prefers, ∅, or $\tilde{K}_i'$), and these may differ from experiment to experiment. *But*, it could be argued, *this is irrelevant, because we know that they correspond to E being true, or E being false.*[47]

Agreed …, but what does this mean exactly? The last sentence, so simple, clear and straightforward, is admirably suited to an hypothesis which is equally simple and clear; the hypothesis which assumes that one of the many experiments has been taken as the *definition* of E (i.e., if the one chosen is K_{13}, then E means E_{13}, or, better, $E_{13}|K_{13}$), and that this experiment is always possible and always successful. It follows that the statement that E is true because a different experiment K_4 has given the signal E_4, can be strengthened by the remark 'since it is certain that the answer E_{13} – that is E – will be obtained if we perform experiment K_{13}, it is unnecessary to do so, because of the trouble, expense and so on; but, if you do not believe it, try it, and you will see!'. In this way, for example, if I derive the height of a distant tower by trigonometric methods, or by observing how long it takes for stones dropped from the top to reach the ground, I can say to someone who does not believe it 'go up and measure it'. And one might allow that the argument is considered in general to be valid, even when the invitation becomes rather less realistic (distance from the centre of the earth, distance between two stars, or two galaxies). But what if someone does not believe it?

In the previous hypothetical case, there was a criterion which could appropriately be assumed as a definition, both because of its meaning and because it could always be applied (at least in principle) with a guarantee of success. In particular, it could be applied to statements about things that are not directly observable and possess disconcerting properties (as in the case of 'waves' and 'particles'). What do we do in the absence of such a criterion? We could define all the experiments K_i (and the simple ones will suffice, it is not necessary to consider strategies) by means of the respective answering signals and observe that, in any case, no matter which K_i are applied, and no matter how many of them, in all successful cases they give a concordant answer; either always E_i, or always $\tilde{E}_i$. This, practically speaking, assures the meaning and uniqueness of the notion related to E (or, more generally, to a quantity), provided that the coincidence of answers for any two methods, K_i and K_j, could be verified experimentally by applying both of them in precisely that same situation (or at least indirectly, by means of a chain of equivalences, each link of which could be verified experimentally).

47 See the discussion of H. Jeffreys that we have already quoted (Chapter 11, the end of 11.1.1).

But what if we come across cases where it is not possible to perform more than one experiment in a given situation? Any statement of the form 'having observed the outcome E_i of the experiment K_i, we know that *had we performed* the experiment K_j we *would have* obtained E_j', is entirely without content, since the assumption is false. It is the same situation as the one we already considered, albeit light-heartedly, at the beginning of Chapter 4 (Section 4.1), when we asked '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'.

It might help to place one's confidence in a more speculative kind of generalization, such as one makes in passing from the direct verification of indirect measurements of length on the 'human' scale to admitting the same thing for inaccessible distances. The generalization required would have to admit the coherence and validity (justified by numerous indirect proofs) of the entire set of concepts, arguments and calculations which constitute the scientific view of the world.

It is a fact, however, that, so far as the 'ordinary man in the street' is concerned, the only reason he believes in these things is a lack of appreciation of the fact that they are more abstruse and delicate than he imagines. The logical situation for him is, under the worst assumptions, the following: he is given the explanation that the fact of E being true (to fix ideas, think of E as an event on which he would like to bet) can be verified in one and only one of the many possible ways provided by performing an experiment K_i and receiving a corresponding answering signal E_i (the choice being made by the experimenter; in any case, *this tells him about the same fact*). But this leading statement conveys nothing to the man in the street, who has no idea of what 'fact' it is that one is dealing with. (For the scientist, too, it is very much an intellectual conviction; but we are not interested in him.) The man in the street only knows, naturally enough, that he might bet on the outcome of an experiment whose choice is in the hands of his opponent. He may therefore think (in theory, of course – not in practice, because it is not nice to be suspicious) that the choice will be made to his disadvantage: for example he may end up trying to draw a white ball from an urn containing only black balls, this being one of the possible choices open to his opponent.

Leaving these more or less picturesque illustrations aside, it would seem that the conclusion – a negative and disturbing one – cannot be other than the following: *one does not succeed in giving an operational meaning to a statement E (or to a quantity X) by means of a collection of statements $E_i|H_i$, which do have operational meaning, without introducing the statement that all the obtainable results E_i are necessarily conformable* (*and this does not have an operational meaning if the H_i are incompatible*).

14 From 'Possibility' to 'Probability'

The logic of certainty only distinguishes events which are either true or false, and which can only be *possible* (uncertain) rather than certain or impossible[48] (for us, in our more or less temporary state of ignorance). We have discussed questions of a critical nature

48 We do not have to worry about 'indeterminacy', considering it (see Sections 9–12) reducible to the case of two-valued logic by means of conditional events.

by remaining within this ambit as a preliminary exploration of the field into which we have to introduce and apply the logic of the probable.

The time has now come to deal with the main critical questions, which specifically concern the subject of direct interest to us: the theory of probability.

We do not want to repeat ourselves by going back to the beginning and starting from scratch: we have already dealt with many questions in the text, and many comments were necessarily made as we developed our approach. It will be more appropriate to refer back to these, to draw the threads together, and to go more deeply into them, in order to provide a synthesis and, finally, what will, hopefully, turn out to be a sufficiently integrated view of the entire subject.

Let us begin by sketching a broad outline, including both those topics for which we shall limit ourselves to recalling our previous remarks – or just adding the odd word here and there – and those which we shall take up again later because they require further analysis or more thorough discussion.

Without further comment, we shall take as axioms those already established as the basis of our subjectivistic formulation. This will in no way prejudice the (technically neutral) possibility of comparison; our starting point, in fact, makes comparison easier, because it represents the minimal set of conditions common to all formulations. The subjectivistic formulation, as we have said repeatedly, is, in fact (and deliberately so), the *weakest* one; its only requirement is coherence, and in no way does it seek to interfere with an individual's freedom to make an evaluation by entering into the merits of it on some other grounds.

In discussing these concepts, we shall provide a comparison with other points of view, which differ in various respects (in the interpretation of the notion of probability, the mathematical details and the qualitative formulation).

Those interpretations of the notion of probability in a (would-be) objective sense that are based on symmetry (the classical conception; equally likely cases), or on frequency (the statistical conception; repeated trials of a phenomenon), provide criteria which are also accepted and applied by subjectivists (as, to a considerable extent, in this book). It is not a question of rejecting them, or of doing without them; the difference lies in showing explicitly how they always need to be integrated into a subjective judgment and how they turn out to be (more or less directly) applicable in particular situations. If one, instead, attempts to force this one or that one into the definitions, or into the axioms, one obtains a distorted, one-sided, hybrid structure.

The mathematical details remain those that derive from the positions we adopted concerning zero probability, countable additivity and the interpretation of asymptotic laws (points which we have already encountered, and commented on, many times). In this regard, we shall have to consider many further points, which we glossed over in Chapters 3, 4 and 6, in order not to overcomplicate the exposition (prematurely), and to add some details concerning a number of new features. These considerations, together with some others, will enable us to sort out, and comment upon, the differences between the axiom system we have adopted here, and that given by Kolmogorov (1933), the formulation which, broadly speaking, has been adopted by most treatments of the last few decades.

Finally, under the heading of 'qualitative formulations', we will have to mention two separate topics. The first concerns the possibility of starting from purely qualitative axioms – that is in terms of comparisons between probabilities of events (this one is

more probable than that one etc.) – without introducing numerical probabilities, but eventually arriving at them by means of comparisons of this kind. The second deals with the thesis that several authors have recently put forward, namely that probabilities are intrinsically indeterminate. The idea is that instead of a uniquely determined value p one should give bounds (upper and lower values, p' and p''). That an evaluation of probability often appears to us more or less vague cannot be denied; it seems even more imprecise, however (as well as being devoid of any real meaning), to specify the limits of this uncertainty.

15 The First and Second Axioms

The entire treatment that we have given was based on a small number of properties, which were justified in the appropriate place in the text as conditions of coherence. In order to develop the theory in an abstract manner, it will now suffice to assume these same properties as axioms.

There will be two axioms (the first and the second) dealing with previsions and a third dealing with conditional previsions. The third one – which is needed in order to extend the validity of the first two to a special case – will be dealt with later (Section 16); we concentrate for the time being on the first two.

Axiom 1 *Non-negativity*: if we *certainly* have $X \geqslant 0$, we must have $\mathbf{P}(X) \geqslant 0$.

Axiom 2 *Additivity* (finite):

$$\mathbf{P}(X+Y)=\mathbf{P}(X)+\mathbf{P}(Y).$$

From these it also follows that

$$\mathbf{P}(aX)=a\mathbf{P}(X), \quad \inf X \leqslant \mathbf{P}(X) \leqslant \sup X,$$

as well as the (Convexity) condition, which includes Axioms 1 and 2:

(*C*) *any linear equation* (*or inequality*) *between random quantities* X_i *must be satisfied by the respective previsions* $\mathbf{P}(X_i)$; in other words,

$$\text{if we } \textit{certainly} \text{ have } c_1X_1+c_2X_2+\ldots+c_nX_n=c \left(\text{or} \geqslant c\right)$$

$$\text{then } \textit{necessarily } c_1\mathbf{P}(X_1)+c_2\mathbf{P}(X_2)+\ldots+c_n\mathbf{P}(X_n)=c \left(\text{or} \geqslant c\right).$$

By taking differences, (*C*) can be written in an alternative form:

(*C'*) *No linear combination of* (*fair!*) *random quantities can be uniformly positive*; in other words, the $\mathbf{P}(X_h)$ must be chosen in such a way that whatever be the given $c_1, c_2,\ldots, c_n$, *there does not exist a* $c > 0$ *such that*

$$c_1\left(X_1-\mathbf{P}(X_1)\right)+c_2\left(X_2-\mathbf{P}(X_2)\right)+\ldots+c_n\left(X_n-\mathbf{P}(X_n)\right) \geq c$$

certainly holds.

We could put forward as a further (possible) axiom one which consists in excluding the addition of other axioms; that is one which considers *admissible*, as prevision-functions $\mathbf{P}$,

all those satisfying Axioms 1 and 2, or equivalently, condition (*C*).[49] On the other hand, this is implicit, since nothing is said to the contrary. In any case, we shall say that every function **P** satisfying Axioms 1 and 2 is *coherent*.

As we have already mentioned (Chapter 3, 3.10.7), a coherent function **P**, defined on some given set of random quantities X (an arbitrary set, in general infinite), can always be extended, preserving coherence, to any other random quantity, X_0, say. From any inequality of the form (*C′*), one can obtain, by solving it with respect to one of the summands (let us assume $c_0 = \pm 1$ and take it to be the one corresponding to X_0; were this not the case, it suffices to divide through by $|c_0|$), an inequality for $\mathbf{P}(X_0)$ of the form

$$\mathbf{P}(X_0) \leqslant \inf\left\{ X_0 + \sum_{h=1}^{n} c_h \left(X_h - \mathbf{P}(X_h) \right) - c \right\} \qquad \left(\text{or} \geqslant \sup\{\ldots\} \right).$$

As a result, we obtain $x' \leqslant \mathbf{P}(X_0) \leqslant x''$, where x' denotes the greatest lower bound and x'' the least upper bound. If $x' = x''$, the extension will turn out to be uniquely defined;

$$\mathbf{P}(X_0) = x' = x'',$$

that is $\mathbf{P}(X_0)$ will be determined by the values given over X. If $x' < x''$, the admissible values for $\mathbf{P}(X_0)$ will consist of all those in a closed interval (as is obvious by convexity). The extension would be impossible if $x' > x''$, but this is ruled out by the observation that there would then exist a linear combination $X_0 + \sum_i c_i \left(X_i - \mathbf{P}(X_i) \right)$ always $> x'$, and another one

$$X_0 + \sum_j c_j \left(X_j - \mathbf{P}(X_j) \right)$$

always $< x''$; their difference ($\Sigma_i - \Sigma_j$; X_0 cancels out) would then turn out to be $> x' - x'' > 0$. But this would mean that there was a contradiction of (*C′*) already contained in X, contrary to the hypothesis.

It follows immediately from this that one can always define a $\mathbf{P}(X)$ for all the X belonging to an arbitrary set of random quantities (in particular, one can always define a $\mathbf{P}(E)$ for every event in an arbitrary collection of events – for example those corresponding to all subsets of a given space), even assuming $\mathbf{P}(X)$ as already assigned in some given field, and extending it. It is sufficient, as we have done here, to carry out the extension for new X one at a time, by means of *transfinite induction* (assuming, of course, the Zermelo Postulate, in order to well-order the X_h; the indices, h etc., will be transfinite ordinals). One has to be a little careful that nothing goes wrong for the X_k which have no

49 Note that we are not dealing here with the basic issue of whether under given circumstances all the coherent evaluations **P** are admissible (subjectivistic conception), or whether only one of them corresponds to reality (objectivistic conceptions). For the objectivist also, it is a question of knowing which **P** are formally admissible (e.g. the **P** which he can adopt when he has the information he is now lacking – about composition of urns, frequency of statistical phenomena etc.), or even that he judges to be possible with respect to the abstract scheme without knowing which concrete events are represented by the symbols E_1, E_2 etc. On the other hand, this is the attitude adopted by the supporters of all points of view when they are faced with the notion of '(abstract) probability space'.

antecedent (such as X_ω, where ω denotes, as usual, the first ordinal which comes after the natural numbers).[50] In our case, however, the contradiction would derive from the comparison between two *finite* linear combinations and should have occurred at the last of the steps corresponding to the X_h which appear (and the fact that there are an infinite number of steps between this X_h and our X_k does not enter into the argument).

Let us return now to the problem of the extension, in order to consider when it turns out to be uniquely defined. One obvious case is that of a random quantity X_0 linearly dependent on those of the original field X; that is belonging to the linear space L generated by the X belonging to X. In this case, the uniqueness of the extension holds for any **P**.

Condition (*C*), however, reveals what the situation is in terms of a particular **P**. Instead of linear relations, we have, in general, linear inequalities, $\sum_i c_i X_i \geq c$, which, solved in terms of X_0 (as above for $X_0 - \mathbf{P}(X_0)$), give random quantities X' and X'', linear combinations of random quantities belonging to the field X (and hence belonging to L), which bound X from below and above: $X' \leqslant X$ and $X \leqslant X''$, respectively. We observe that the problem is the same one that we already encountered in a special case (Chapter 3, 3.12.4), and by passing, as here, to the general and abstract case, we also reached essentially the same conclusions. As we vary **P** (defined over X, and hence on L, and, in particular, for X' and X''), the X' for which $\mathbf{P}(X')$ is a maximum, $\mathbf{P}(X') = x'$, will also vary (or, if x' is an upper bound rather than a maximum, the X' to be chosen in order to obtain $\mathbf{P}(X')$ arbitrarily close to x' will vary): similarly for X''. Having chosen X and X'' in this way, we have $X' \leqslant X \leqslant X''$, with $\mathbf{P}(X'' - X') = x'' - x'$ (or $x'' - x' + \varepsilon$, with $\varepsilon > 0$ arbitrary, in the case when they are not the maximum and minimum). In general, therefore, one has a uniquely defined extension if upper and lower bounds of X exist for which the difference $\Delta = X'' - X' \geqslant 0$ has prevision $\mathbf{P}(\Delta) = 0$, or such that $\mathbf{P}(\Delta) < \varepsilon$ (for arbitrary, fixed $\varepsilon > 0$).

In order to denote what can be said about the probability (or prevision) outside some given linear space L in terms of the prevision function **P** defined over it by the evaluations of probability (or prevision), it is convenient to use the same notation (*mutatis mutandis*) as we used in Chapter 6, 6.4.4.

We thus denote by

$$\mathrm{P}_{\mathrm{L}}^{-}(X) = x', \quad \mathrm{P}_{\mathrm{L}}^{+}(X) = x'', \quad \mathrm{P}_{\mathrm{L}}^{\pm}(X) = x,$$

the minimum and maximum (previously indicated in the text by x' and x'') of the values $\mathbf{P}(X)$ which are compatible with the knowledge of **P** over L, and, respectively, their common value (if they are equal).

We shall say more about this – for other reasons – in Section 19.4.

---

50 Lebesgue measure, too, can be extended, preserving countable additivity, to an arbitrary non-measurable set, and hence to an arbitrary number of such sets, one at a time. In this case, however, an infinite number of steps can lead to a contradiction without any single step doing so (in the same way as a convergent series remains such if we replace the 1st, 2nd, 3rd, …, terms with 1, and so on for any finite number, but not if we replace an infinite number of terms).

16 The Third Axiom

Conditional probabilities $\mathbf{P}(E|H)$, or conditional previsions, $\mathbf{P}(X|H)$, are expressible, in cases where H has nonzero probability, in terms of the unconditional probabilities by means of a formula which, in an abstract, axiomatic treatment, can be taken as a *definition*:

$$\mathbf{P}(E \mid H) = \mathbf{P}(EH) / \mathbf{P}(H), \qquad \mathbf{P}(X \mid H) = \mathbf{P}(XH) / \mathbf{P}(H).$$

In this case, there is nothing much to add, apart from noting that here, too, an extension (in the sense of $\mathbf{P}_{\mathrm{L}}$) gives rise to an interval of indeterminacy:

$$\mathbf{P}_{\mathrm{L}}^{-}(X \mid H) \leqslant \mathbf{P}(X \mid H) \leqslant \mathbf{P}_{\mathrm{L}}^{+}(X \mid H).$$

To see this, suppose that $\mathbf{P}_1$ and $\mathbf{P}_2$ are two extensions of $\mathbf{P}$ as given over L, and that these give to XH and H the values

$$\mathbf{P}_1(XH) = x_1, \quad \mathbf{P}_2(XH) = x_2, \quad \mathbf{P}_1(H) = h_1, \quad \mathbf{P}_2(H) = h_2.$$

In addition to $\mathbf{P}_1$ and $\mathbf{P}_2$, their convex combinations,

$$\mathbf{P}_\lambda = \lambda \mathbf{P}_1 + (1-\lambda)\mathbf{P}_2 = \mathbf{P}_2 + \lambda(\mathbf{P}_1 - \mathbf{P}_2) \ (0 \leqslant \lambda \leqslant 1),$$

will also be extensions of $\mathbf{P}$, and will give

$$\mathbf{P}_\lambda(X \mid H) = \mathbf{P}_\lambda(XH) / \mathbf{P}_\lambda(H) = \frac{x_2 + \lambda(x_1 - x_2)}{h_2 + \lambda(h_1 - h_2)}.$$

Since the denominator does not vanish (for $0 \leqslant \lambda \leqslant 1$; or at most at one of the end-points if one of the h_i is zero, a case that we shall not consider now, however), the hyperbola increases or decreases monotonically between the extreme values

$$\mathbf{P}_1(X \mid H) = x_1 / h_1 \quad \text{and} \quad \mathbf{P}_2(X \mid H) = x_2 / h_2.$$

In the extension, the set of possible values for $\mathbf{P}(X|H)$ is thus an interval as asserted.

If $\mathbf{P}(H) = 0$, we have a new situation. Does it make sense to consider this case? And, if so, for what purpose? If one were to take the formula, with $\mathbf{P}(H)$ in the denominator, as the actual, unique definition of conditional probability and prevision, then the concept, in this case, would become meaningless. If the meaning were to be assigned in some other, direct, way – for example (as was done in Chapter 4, in line with the subjectivistic point of view), by means of conditional bets – then the meaning would be retained.

But the theorem which expresses coherence, connecting it to the nonconditional $\mathbf{P}$ (the theorem of compound probabilities), no longer holds (and neither does the criterion of coherence) if its formulation (Chapter 4, Section 4.2) has to be in terms of the existence of a 'certainly smaller' loss. In order to extend the notions and rules of the calculus of probability to this new case, it is necessary to strengthen the condition of coherence by saying that *the evaluations conditional on H must turn out to be coherent conditional on H* (i.e. under the hypothesis that H turns out to be true). This is automatic if one

evaluates $\mathbf{P}(H) \neq 0$, in which case we reduce to the certainty of a loss in the case of incoherence. The loss for $\tilde{H}$ (Chapter 4, Section 4.3) is, in fact, the sum of the squares of $\mathbf{P}(H)$ and $\mathbf{P}(EH)$; but if $\mathbf{P}(H)$, and therefore $\mathbf{P}(EH)$, are zero, this loss is also zero in the case $\tilde{H}$ (which has probability = 1, and is, in any case, possible).

Although this strengthening of the condition of coherence might seem obvious, we had better be careful with it. There are several other forms of strengthening of conditions, often considered as 'obvious', which have consequences that lead us to regard them as inadmissible. In this case, however, there do not seem to be any drawbacks of this kind; moreover, the 'nature' of the strengthening of the condition seems more firmly based on fundamental arguments (rather than for conventional or formal reasons, or for 'mathematical convenience') than others we have come across, and to which we shall return later. In any case, we propose to accept the given extension of the notion of coherence, and to base upon it the theory of conditional probability, without excluding, or treating as special in any way, the case in which one makes the evaluation $\mathbf{P}(H) = 0$.

If we wish to base ourselves upon a new axiom, we could express it in the following way:

Axiom 3 *The conditions of coherence* (*Axioms 1 and 2*) *must be satisfied, also, by the* $\mathbf{P}_H$ *conditional on a possible H*, where

$$\mathbf{P}_H(E) = \mathbf{P}(E \mid H), \qquad \mathbf{P}_H(E \mid A) = \mathbf{P}(E \mid AH)$$

is to be understood.

This means that $\mathbf{P}_H$ is the prevision function that we may have ready for the case in which H turns out to be true, and the axioms oblige us to make this possible evaluation in such a way that if it is to have any effect it must be coherent. This is implicit in the previous definition if one makes the evaluation $\mathbf{P}(H) \neq 0$. Axiom 3 obliges us to behave in the same way, simply on the grounds that H is possible and we might find ourselves actually having to behave according to the choice of $\mathbf{P}_H$ – even if, in the case in which we attribute probability 0 to the hypothesis H, the sanction provided by the losses does not apply outside of the case $\tilde{H}$.

Axiom 3 permits us to define the ratio of the probabilities of two arbitrary events – even if they have zero probabilities – in the manner already introduced in Chapter 4, 4.18.2. In Section 18.3, we shall expressly return to the topics concerning zero probabilities, topics previously dealt with in Chapter 4, 4.18.3–4.18.4.

17 Connections with Aspects of the Interpretations

The axioms of an abstract theory are, as such, arbitrary and independent of this or that interpretation (at this level, interpretations do not, strictly speaking, exist; or, to put it a little less strongly, one might say that they are ignored).

It goes without saying, however, that the choice of axioms is influenced by the interpretation they will have when the theory is applied in the field for which it has, in fact, been constructed, and on which one would like it to turn out to be adequately modelled.[51]

---

51 As someone rather neatly put it – Frechet attributes the remark to Destouches – a book which starts off with axioms should be preceded by another volume, explaining how and why these axioms have been chosen, and with what end in view.

In the case of the theory of probability, any judgement about the adequacy of the axioms depends on one's concept of probability and, in addition to the subjectivistic concept, which we have adhered to throughout, we shall also have to consider the 'classical' and 'statistical' concepts.

From the subjectivistic point of view, the axioms are valid in that they are a translation of the necessary and sufficient conditions for coherence (our starting point in Chapters 3 and 4). It follows that no other axioms can be admitted (since these would introduce further restrictions).

Mention should be made of a formulation which is subjectivistic in a purely psychological sense, and in which no axioms would be acceptable. This is the approach in which one simply thinks of evaluations of probability – in general, incoherent – being made by some, arbitrary, individual. It is clear that without sufficient preparation and thought everyone would give incoherent answers in every field (e.g. by estimating distances, areas, speeds etc. in an incoherent manner). This does not imply, however, that there exists – albeit only in the individual's own mind – a different theory (e.g. a non-Euclidean geometry) to be made an object of study. The object of study could only be the extent of his intuitive inability to understand the conditions of coherence, and to avoid breaking them. Otherwise, one would have to say that, in a system of bets, he deliberately chooses to behave in such a way as to lose.

From the classical point of view – probability 'defined' as the ratio of favourable cases to possible cases, all considered 'objectively' equally likely for reasons of symmetry – the axioms are true by virtue of the laws of arithmetic (sums of fractions, together with certain other details which are required to achieve the necessary rigour). There is, for any given application, just one admissible **P**. It would appear to be valid to consider infinite partitions into equally probable cases (by virtue of symmetry).

As an extension of this point of view, one might consider the 'necessary' conception, which takes probabilities of a collection of events, possibly outside the range of cases considered in the 'classical' approach, to be uniquely defined for *logical* reasons. A typical example – one that accepts the possibility of 'an infinite number of equally likely cases' – is provided by Jeffreys' admission of improper initial distributions (e.g. uniform in X, or in $\log X$ etc.). It appears that Carnap's point of view is similar to this.[52]

From the statistical point of view – where probability is regarded as 'idealized frequency'[53] – additivity always holds for arithmetic reasons (as in the classical case). According to this conception (again as in the classical case), there should be a unique admissible **P**. It is difficult to attempt to venture hypotheses about the interpretation of more delicate cases (e.g. zero probability).

An attempt to make the statistical conception more precise consists in defining probability not as an 'idealization', but rather as a *limit* of the frequency (as the number of trials, n, tends to ∞). In throwing a die, the limit-frequency of 'evens' is without doubt the sum of the limit-frequencies of '2', '4' and '6', if these limits exist (and this is assumed

52 It is always difficult to judge whether similarities are real or apparent (particularly between authors with different backgrounds, working in different fields).

53 This phrase does not really convey anything, but it is the only way to refer to the many confused explanations given by the supporters of this conception, and it may be that, in fact, there is nothing of substance to 'understand' (alternatively, it may be me who lacks the resources necessary for success in this toilsome venture).

to be the case in the scheme we are considering). It seems equally clear that such additivity does not (necessarily) hold for infinite partitions: a 'die with an infinite (countable) number of faces' could well turn up each face with limit-frequency zero[54] (indeed, *it should* do so, if we continue to admit, in some shape or form, the assumption of equal probabilities for all the infinite faces). But this seems to get overlooked (see Chapter 3, 3.11.6, the case $C9 = N9$).

Although it is really going beyond our aim of giving a critical analysis of particular attempts at axiomatization, it seems necessary to spend some time on the following case in order to make some comments (we shall refer to the version given by von Mises, which is the most developed, and which was in favour for a time).

On the one hand, it appears that the hypothesized sequences with well-determined limit-frequencies should represent 'idealizations' of problems of 'repeated trials'. This would seem to be so in view of introductory remarks alluding to ideas like the 'empirical law of averages' and because of an additional restriction ('Regellosigkeitsaxiom', or the axiom of 'nonregularity'), which is intended as a summary, in objective and descriptive terms, of the apparent effects of the *independence* of successive trials. Should one exclude periodicities? The grouping of the results in blocks (e.g. each 'colour' at least three times in a row)? The sequences definable in terms of simple mathematical formulae, or sentences not exceeding 100 words? On each occasion one would probably answer yes; but in actual fact there is never any reason to call a halt before having excluded all the possibilities, nor, conversely, any justification for absolutely excluding any given case.

On the other hand, if one wishes to consider the actual case of a sequence of trials (in general unlimited, but to begin with let us assume it to be limited to a finite number, n), independent and with equal probabilities – according to the concepts derived from such a formulation – one must not think in terms of the 'sequences' of the previous model. One must think of n parallel sequences; that is a sequence of n-tuples representing (let us say) a fictitious infinity of 'copies' of the actual sequence of n trials (for $n = \infty$, a fictitious infinity of copies of the whole actual sequence). Only in this absurd supermodel do the simple and obvious ideas of independence and equal probabilities make any sense, and is one able to (correctly) conclude that in the actual (Bernoulli) sequence one has stochastic convergence (weak, strong, in mean-square), but not definite convergence (as was postulated in the original scheme).

---

54 A 'reasonable enough' example might be obtained by saying that each face of the die has probability $p_h = e^{-1}/h!$ of occurring h times (in an infinite series of trials). Alternatively, if one prefers, one could say that out of 10 000 faces occurring 'at random' we will have 'in prevision' (or, less appropriately, 'on average')

10 000 p_h =	3679	3679	1839	613	153	31	5	1
where h =	0	1	2	3	4	5	6	7

(i.e., in the infinite series of trials 0, 1, 2,…, 7 repeats will occur, but 8 or more will not occur even once – on average about 0·10 times).

This is the Poisson distribution with $m = 1$, which holds asymptotically for the game of matching n objects ($n \to \infty$), or for that of n drawings with n balls (e.g. 90 drawings, with replacement, of the 90 numbers in bingo), again as $n \to \infty$. We have remarked that this example is 'reasonable enough', but it is no more than this, because the choice of this scheme from among infinitely many others is arbitrary. One could, for example, vary the scheme for drawing the balls, by assuming that as n, the number of balls, increases, the number of drawings is not n, but $2n$, or n, or $\sqrt{n}$ etc.

The original scheme is therefore a sham. For the purpose of winning over the unwary – who do not notice the sleight of hand – properties are attributed to it which look like the probable properties of actual sequences of 'repeated trials', but which are, in fact, misleading and incompatible with them. By mysterious manipulations of an infinite number of such shams, one finally succeeds in saying those things which could have been said directly anyway (i.e. that the trials are independent and equally probable). The fruits of these labours are that one now does not understand the (subjective) meaning of the words, and that the flurry of sophistical acrobatics has created the illusion that one has established or produced an 'objective' something or other.

In bringing to a close our summary of the various interpretations, we repeat that the subjectivistic conception is not in opposition to any of them, but rather that it utilizes all of them. It is simply a question of rejecting the claims of exclusiveness that lead to incomplete and one-sided theories, of correcting the distortions made in order to make them appear objectivistic, of considering them as methods whose appropriateness varies with the situation, and of seeing them as having one and the same function: that of aiding the individual in his task of evaluating the probabilities (always subjective) to be attributed to events of interest.

18 Questions Concerning the Mathematical Aspects

18.1. We now turn to an examination of those aspects of a purely mathematical or formal nature. In a certain sense, we will be considering the properties of the function **P** and the meaning of the implications of these properties. We refer here to meaning in a formal sense; without reference – except incidentally, and for purposes of clarification – to the different interpretations and assumptions which precede the choice of axioms (concerning which see Section 17).

In order to provide an overall perspective, it will be convenient to present the various questions – including those we have already dealt with – in the context of a comparison with the Kolmogorov axiom system,[55] a formulation which is well known to everyone.

The basic differences are:

i) we REJECT the idea of 'atomic events', and hence the systematic interpretation of events as sets; we REJECT a ready-made field of events (a Procrustean bed!), which imposes constraints on us; we REJECT any kind of restriction (such as, for example, that the events one can consider at some given moment, or in some given problem, should form a field);
ii) we REJECT the idea of a unique **P**, attached once and for all to the field of events under consideration; instead, one should characterize *all* the admissible **P** (the set P); P turns out to be closed (Chapter 3, Section 3.13), and thus any **P** adherent to P in fact belongs to it (a property which does not hold in the Kolmogorov system);

55 A. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin (1933). The first time I developed a systematic discussion in the context of a comparison with this theory was in 'Sull'impostazione assiomatica del calcolo delle probabilità', *Annali Triestini*, **XIX**, University of Trieste (1949).

iii) our approach deals directly with *random quantities* and *linear* operations upon them (events being included as a special case); we thus avoid the complications which arise when one deals with the less convenient Boolean operations;
iv) we REJECT countable additivity (i.e. σ-additivity);[56]
v) we REJECT the transformation of the theorem of compound probabilities into a definition of conditional probability, and we also REJECT the latter being made conditional on the assumption that $\mathbf{P}(H) \neq 0$; by virtue of the exclusions we have made in (iv) and (v), the construction of a complete theory of zero probability becomes possible;
vi) Kolmogorov's proof of the compatibility of his axioms is open to criticism (see the paper of mine quoted in the footnote at the beginning of this section); this is, however, a problem that can be resolved, and it has no substantive implications.

To a greater or lesser extent, all these matters have been touched upon already, either in the text or in this Appendix. We shall only concern ourselves now with those aspects which require further analysis or more detailed discussion.

18.2. *Zero probabilities*. Let us first of all go back to our earlier discussion (at the end of Section 16), and let us repeat the proofs and definitions that we gave (in Chapter 4, 4.18.2), basing ourselves now on Axiom 3.

Axiom 3 permits us to define the ratio of the probabilities of two arbitrary events, A and B, by observing that for all H which contain A and B (i.e. $H \supset A \vee B$) the ratio $\mathbf{P}(A|H)/\mathbf{P}(B|H)$ does not change (except possibly to become indeterminate, 0/0). Suppose, in fact, that H' and H'' are events containing $A \vee B$, and that they do not give rise to the case of 0/0, and let $H = H'H''$ be their product, which also contains A and B (or one could take H to be $A \vee B$). Since $\mathbf{P}_{H'}$, and $\mathbf{P}_{H''}$ must be coherent, we can write

$$\mathbf{P}_{H'}(A) = \mathbf{P}_{H'}(AH) = \mathbf{P}_{H'}(H).\mathbf{P}_{H'}(A \mid H).$$

But $\mathbf{P}_{H'}(A \mid H) = \mathbf{P}(A \mid HH') = \mathbf{P}(A \mid H)$, because $H \subset H'$, $HH' = H$. Finally, we have $\mathbf{P}_{H'}(A) = \mathbf{P}_{H'}(H).\mathbf{P}(v)$, and $\mathbf{P}_{H'}(B) = \mathbf{P}_{H'}(H).\mathbf{P}(B \mid H)$, and hence it follows that

$$\frac{\mathbf{P}_{H'}(A)}{\mathbf{P}_{H'}(B)} = \frac{\mathbf{P}_{H'}(H).\mathbf{P}(A \mid H)}{\mathbf{P}_{H'}(H).\mathbf{P}(B \mid H)} = \frac{\mathbf{P}(A \mid H)}{\mathbf{P}(B \mid H)}.$$

The same holds true for every H'', and, finally, in order to obtain the ratio, it suffices simply to take $H = A \vee B = A + B - AB$; in this case we certainly have $\mathbf{P}_H(A) + \mathbf{P}_H(B) \geqslant 1$ and 0/0 cannot occur.

In this way, the formula $\mathbf{P}(E|H) = \mathbf{P}(EH)/\mathbf{P}(H)$ *is always meaningful and valid*, and the same is true for every application of the theorem of compound probabilities, and, more generally, for any operation involving probability ratios, so long as they make

56 It is worth mentioning, incidentally, that if one decides to proceed in the direction of assuming things for 'mathematical convenience', then not even countable additivity appears to be sufficiently restrictive. Several authors, including Kolmogorov himself, have recently proposed axioms ('perfect' additivity, and such-like) that make the principles of probabilistic reasoning, essential to every human being, completely dependent on the abstruse subtleties of set theory at its most profound. See, for example, D. Blackwell, 'On a class of probability spaces', in *Proc. 3rd Berkeley Symp*., II, pp. 1–6 (1956), and other works referred to therein.

sense (i.e. so long as one does not introduce the nonsensical, indeterminate expressions, $0/0$, $0/\infty$, ∞/∞, which must be avoided by means of the procedure used for defining the ratio in each case).

There are, therefore, different orders, or layers, of zero probability (as we have already noted in Chapter 4, 4.18.3, where we also saw how very rich and complicated structures of such layers could be constructed). We shall see later, in 18.3, what the situation would necessarily be in this respect were we to assume the axiom of countable additivity (or, at least, if the condition were assumed to hold in some specific example or other). For the present, however, let us return to the general case.

The theorem of total probability will have to be interpreted in the following extended sense (which includes the case of zero probabilities): given n incompatible events, E_1, E_2,..., E_n, the probability of the sum-event E is the sum of the nonzero probabilities, if there exist any, and, if not, it is *the sum of the zero probabilities of maximal order*. If, for example, E_3 is of maximal order (i.e. $\mathbf{P}(E_h)/\mathbf{P}(E_3) < \infty$, $h = 1, 2,..., n$), the sum-event has probability $\mathbf{P}(E) = \mathbf{P}(E_3)$ if for all $h \neq 3$ the preceding ratio is not only $<\infty$ but in fact $=0$. In general, it is given by

$$\mathbf{P}(E) = \mathbf{P}(E_3)\sum_{h=1}^{n} c_h,$$

where $c_h = \mathbf{P}(E_h)/\mathbf{P}(E_3)$ ($c_3 = 1$; the other c_h may be zero, in which case they do not count, or they may be greater than, or less than, 1).

The introduction of conditional probability freed from the restriction that the 'hypothesis' have nonzero probability, and the consequent possibility of comparing zero probabilities, is important both from a conceptual and from a practical point of view. This importance derives not so much from the fact that we can see the potential usefulness in interesting applications, but rather from the warning it provides against inaccurate ways of approaching – or, at least, of expressing – certain questions. We have in mind methods of approach that either lead to confusion or to an over-hasty choice of the path and interpretation to be followed, because of the absence of a precise meaning to which one can refer.

No doubt some will regard discussion of this kind as rather artificial and academic; nothing more than hair-splitting *ad infinitum*. They may be right and they will do well to pose their problems in such a way as to avoid the difficulties. But in order to do this, they must first be able to recognize the difficulties as such, so as to overcome them without lapsing into naïvety or contradiction. In any case, since there do exist differences of opinion in this respect, and since the one which I consider to be correct, and which I uphold, differs from that which forms an integral part of the theory currently most in favour, there is no alternative, in the present context, but to consider the matter more deeply.

But why worry about events with zero probability? Are they not, for this very reason, eventualities which can be ignored?

From time to time, someone imagines that he had discovered the way of eliminating the problem altogether, by *establishing* that the values 0 and 1 must be reserved for the probabilities of the impossible event and the certain event, respectively. Every possible event should have positive probability (strictly less than 1). It is easy to see – and we shall do so presently – that this leads in our case to the same kind of absurdities as one encounters when trying to invent a measure which only assigns zero to the empty set. It is only in the very simplest examples (i.e. those where we only meet finite or countable

partitions) that it may happen, *by chance*, that there are no possible events with zero probability, or that, if there are an infinite number of them, their union still has zero probability (a case in which, to some extent, one might regard them as eventualities that can be ignored). If, on the other hand, we considered a nondenumerable partition, we would have to conclude that it was impossible to consider a nondenumerable partition into *possible* cases (because at most a countable number can have positive probability if their sum is not to become infinite $\Sigma p_h = \infty$). We need go no further than the logic of certainty to see the absurdity of this statement.

The major difference between events of zero probability and impossible events is the following: the union of an infinite number of the former can have a nonzero probability (and may even be the certain event), whereas any union of the latter can only be an impossible event.

It is in this setting that one comes across the most controversial question of them all; that of 'countable additivity'. If we limit ourselves to a discussion of it in the context of events having zero probability, countable additivity implies that taking a countable union will never yield an event with positive probability (and certainly not the certain event). One has to examine the specific question of whether it is possible and appropriate to assume this property as an axiom of probability (the property holds, as is well known, for Lebesgue measure, where the nature of the definition excludes the cases for which it would not hold[57]). The majority opinion is that the answer is yes. In my opinion, this is a consequence of external factors, which, generally speaking, are not examined in order to check whether or not they correspond to the essential nature of the problem.

Certain aspects of conditional events also involve us in a consideration of the problems that derive from the presence of events with zero probability. If an event is possible, then – independently of the probability attributed to it, even if it is zero – events conditional on it, and bets related to these events, can always be considered. In this way (by means of the above-mentioned formulae, which we need not consider here), it becomes possible to compare all zero probabilities. They may be of the same order (i.e. having a finite ratio); or of a different order (ratio equal to zero, or, conversely, to infinity). For the purpose of providing an analogy, we have a situation similar to that which arises in comparing two geometrical objects of zero volume; they can be compared by considering the ratios of their areas, or of their lengths, depending on whether they are both two dimensional, or one dimensional. On the other hand, we would say that one was of smaller order if it were a line segment while the other were part of a plane.[58]

We are not concerned with pushing the analogy too far, because the geometrical case has certain special features of its own. What is common to both is the idea of measures of different orders (or, if one prefers, of non-Archimedean quantities). However, the example is to be understood in a purely illustrative sense, with a warning that one should not take into account notions like dimension, distance, volume, limit, cardinality and so on.

---

57 As in the example given by Vitali (quoted in Chapter 6, 6.5.9). See G. Vitali and G. Sansone, *Moderna teoria delle funzioni di variabile reale*, Zanichelli, Bologna (1935), part I, pp. 56 ff.

58 A more systematic method of comparison – for simplicity, we shall always refer to ordinary, three-dimensional space – would be to consider, for each set I, the set of points I_ρ whose distance from I is less than ρ, and the function $V_1(\rho) = \mathrm{Vol}(I_\rho)$ (volume of I_ρ). One can now define the *ratio of the measure of two sets* I' and I'' to be the limit as $\rho \to 0$ of the ratio $V_I'(\rho) / V_I''(\rho)$ (if it exists). It does not always exist, but in the most 'regular' cases we have $V(\rho) = k\rho^{3-d}\,(1 + 0(\rho))$; in other words, $V(\rho)$ is comparable with a power ρ^α $(0 \leqslant \alpha \leqslant 3)$, and, in particular, volumes, areas, lengths, the number of isolated points are given by the coefficients k in the cases for which α turns out to be 0, 1, 2, 3 ($d = 3 - \alpha$ is the number of dimensions, $d = 3, 2, 1, 0$).

Let us just mention that the consideration of probability as a non-Archimedean quantity would permit us to say, if we wished, that 'zero probabilities' are in fact 'infinitely small' (actual infinitesimals) and only that of the impossible event is *zero*. Nothing is really altered by this change in terminology but it might sometimes be useful as a way of overcoming preconceived ideas. It has been said that to assume that

$$0+0+0+\ldots+0+\ldots=1$$

is absurd, whereas, if at all, this would be true if 'actual infinitesimal' were substituted in place of 'zero'. There is nothing to prevent one from expressing things in this way, apart from the fact that it is a useless complication of language, and leads one to puzzle over 'les infiniment petits' [the infinitely small].

Despite all that has been said, some readers may still be of the opinion that all these things are pointless hair-splitting anyway – and, in a certain sense, I would like to reply YES. The fact remains, however, that, paradoxically, the only way of dealing with these things is to think about them and analyse them in detail, carefully studying the most valid and appropriate way of setting them aside, case by case. Even in those cases where approximate answers are preferable to exact ones (because of the illusory nature of the exactness), it is especially important to be doubly precise in one's arguments, in order to know which things remain valid, and which require modification, when the reasons for, and the degree of, this illusory exactness are taken into account.

To this end, we shall make use, among other things, of the ideas that were developed concerning the 'precision' factor, and we shall arrive at conclusions which (hopefully) will appear reasonable, sensible and, perhaps, obvious. But this feeling will only be justified when we have arrived at the conclusion by means of an accurate evaluation of alternative suggestions, which clarifies just what is, and what is not, really significant and well founded.

18.3. *Countable additivity*. An extensive treatment of this topic was given in Chapter 3, Section 3.11, and we have also referred to it on many subsequent occasions. Let us recall the main points as a prelude to making some further critical comments.

The property of additivity, which we have assumed as an axiom, says that in a *finite partition* the sum of the probabilities must equal 1. In other words, if $E_1, E_2,\ldots, E_n$ are exclusive and exhaustive, the probabilities $p_1, p_2,\ldots, p_n$ attributed to them must be non-negative with sum equal to 1. In fact, this is not merely a necessary condition for the evaluation to be coherent and admissible but it is also sufficient.

In the case of an infinite partition into events E_h ($h \in H$, where H is arbitrary), we can only say, on the basis of our axiom, that the sum of every finite number of the p_h must be $\leqslant 1$: in other words, that at most a countable number of them can be positive ($\neq 0$), and that for such values[59] we must have $\Sigma p_h \leqslant 1$. If, in particular, the set of positive p_h has sum = 1, then the E_h with zero probability also have zero probability when taken

59 Even if there are an infinite number of them one can speak of a 'sum' in the sense of 'upper bound of the sum of a finite number of terms' (if one thinks of the 'sum of the series' no conclusion would be legitimate). If we denote by Σ the upper bound (possibly $+\infty$) of the sums of a finite number of terms, we may denote in this way the sum of an arbitrary infinite number of non-negative numbers, and, in particular, of events. For example, $\Sigma E_h(h \in k)$ will denote the number of successes among those E_h for which $h \in K$, and we observe that the standard convention – see Chapter 1, Section 1.9 – in which $(h \in K) = 1$ or $(h \in K) = 0$ according as h belongs, or does not belong, to K, allows one to write $(h \in K)$ as a factor, on the same line, instead of as an index written below the Σ sign (see the application, which follows shortly, with $(p_h = 0)$ for $(h \in K)$ where K = 'the set of the indices for which $p_h = 0$'). If the E_h are incompatible, the sum is necessarily either 0 or 1.

together: that is it turns out that the union $E = \Sigma E_h$ ($p_h = 0$) also has zero probability, $\mathbf{P}(E) = 0$. If, on the other hand, we obtain $\Sigma p_h = P < 1$, that is if a probability $1 - P$ is *missing* in the partition, then the possibilities are as follows: if there are a finite number of events with nonzero probability, then this missing probability is necessarily that of E = the union of the events with zero probability; otherwise, it may be attributed arbitrarily to E and to $\tilde{E}$[46] = the union of the events with positive probability;

$$\mathbf{P}(E) = P', \quad \mathbf{P}(\tilde{E}) = P + P'', \quad P' + P'' = 1 - P.$$

To summarize: given the probabilities p_h of the events E_h of a partition, if their sum is = 1 then the probabilities of all the events depending on it – that is sums of a finite or infinite number of events in the partition – are uniquely determined. In this case, we shall say that the probability $\mathbf{P}(E)$ is countably additive on the partition $\{E_h\}$. Otherwise, this only holds for event-sums of a finite number of the E_h, or for their complements: in any other case, a margin of indeterminacy equal to $1 - P = 1 - \Sigma p_h$ remains;

$$p' = \sum p_h \left(E_h \subset E\right) \leqslant \mathbf{P}(E) \leqslant 1 - \sum p_h \left(E_h \subset \tilde{E}\right) = p'' = p' + (1 - P).$$

Let us be clear that 'indeterminacy' simply means that the extension is not, in general, uniquely defined; one only has bounds. There is no 'indeterminacy' in any specific sense; such as being 'barred' from attributing a well-defined value to $\mathbf{P}(E)$. It is simply that E is not one of the events whose probability has already implicitly been evaluated by virtue of our evaluations for the E_h; it is just one of the many for which our choice is more or less open. We are completely free in our choice (i.e. can give $\mathbf{P}(E)$ any value between 0 and 1) in the particular case in which all the events of the partition have been attributed zero probability (and E is not the sum of a finite number of the E_h, nor the complement of such a sum). This happens in the case of a continuous distribution (on the line, or in the plane, or in ordinary space, ...) for which we have established only that it is 'without concentrated masses', or for a countable number of exclusive (and exhaustive) events of zero probability.

Conclusions of this kind may be hard to accept, or perhaps may even appear paradoxical. At least, the way in which many authors bend over backwards to avoid them – by introducing some new axiom (or 'strengthening' the existing ones) – seems to suggest that this is the case. The following are some of the kinds of restrictions which could be imposed:

(Z) denying that it is legitimate to attribute zero probability to a possible event;
(Za) denying that a union of events with zero probability can have nonzero probability;
(Zb) as (Za), but only considering a countable number of events;
(Ka) assuming countable additivity for arbitrary partitions;
(Kb) as (Ka), but only for countable partitions.

We have introduced the letters Z, Za, Zb, Ka, Kb, in order to facilitate references to these 'axioms'; in what follows, we shall, of course, argue against them.

The first and the last have actually been proposed; Za and Zb are progressively weaker versions of Z and are also special cases of Ka, Kb, respectively; the inclusion of intermediate possibilities would only serve the purpose of pointing out these connections.

First of all, we should draw attention to the lack of any real arguments on the part of those who support such a restriction. It is usually presented as a 'natural' extension of the theorem of total probability as $n \to \infty$ (as, for example, in Cramèr); or as a 'natural' property by analogy with Lebesgue measure (and this is the most common idea); or by Baire extension by continuity (like, for example, in Feller). In other words, for 'mathematical' reasons, and not for reasons relating to probability theory.

One mathematical consequence of this is that it becomes impossible to think of a $\mathbf{P}(E)$ defined for all the events which could be formed on the basis of a nondenumerable partition. An example is provided by the power set of any set whose cardinality is that of the continuum (by virtue of the results of Vitali, Lebesgue, Banach, Kuratowski and Ulam, concerning the impossibility – except in trivial cases of 'concentrated mass' at a finite or countable number of points – of extending σ-additive measures to all the subsets of a nondenumerable set[60]). To admit σ-additivity is to contradict the basic idea that one can attribute to any uncertain event whatsoever a probability – without any, logically inexplicable, discrimination between one event and another. Of course, it could happen that this 'basic idea' itself gives rise to conflicts with other requirements: for example, were it true that there does not always exist an extension of a finitely additive **P**, we should have had to re-examine the whole question of whether, and in what way, a mathematical theory of probability was possible (with goodness knows what weakening of the axioms[61]). The fact that such a disaster does not occur for finite additivity, but does occur if one attempts to replace it with σ-additivity, clearly indicates that the substitution is entirely inappropriate.

If one accepts the subjective concept of probability, the conclusion becomes even more obvious.

In order to reach this conclusion, it was not even necessary, in fact, that the contradiction of not being able to find an arbitrary **P** came to light. It was sufficient that the choice was restricted in a way which appeared to preclude each individual being permitted an unfettered evaluation. And this occurs even in the case of a countable partition. Let us suppose that there are a countable infinity of 'possible cases' and – in order to avoid thinking of points or sets on the real line which appear 'special' in some way – let us imagine that they are represented by points on the circumference of a circle, whose distance apart is a rational multiple of 2π (i.e. by taking the origin at an arbitrary one of these points, with $\theta = 2k\pi$, k rational, $0 \leqslant k \leqslant 1$). That an axiom should not permit me to attribute probabilities which are negative, or have sum greater than one, is something which can be clearly understood as a condition of coherence; it does not impose any

60 From our point of view, it suffices that this has been established for some sets. It appeared preferable, therefore, not to weigh down the text with details of how the result has been proved (Ulam) provided that the cardinality of the set is not 'inaccessible'.

61 A more 'minor' difficulty may serve as an example. A paradox, due to Haussdorff, says that a spherical surface can be divided into three sets, A, B, C such that each is superimposable both on each of the others and on their union; it follows that any 'measure' which was finitely additive and invariant under rigid motions would assign to these sets both $\frac{1}{3}$ and $\frac{1}{2}$ (and $\frac{2}{3}$ and so on; as is well-known, one can logically deduce anything if one starts from something ridiculous). This contradicts geometric intuition, but not the idea of probability (nor the 'axiom of choice'). As Paul Lévy said, in order to refute this interpretation, the simple fact is that 'the continuous in higher dimensions is even more complicated than we thought'. (See E. Borel, *Les paradoxes de l'infini*, Gallimard, Paris (1946); Paul Lévy, 'Les paradoxes de l'infini et le calcul des probabilités', and a note by Borel, in *Bull. Sci. Math* (1948), pp. 184–192.)

restriction on my freedom of opinion. But suppose that an axiom (like *Zb* or *Kb*) prohibits me from attributing the same probability, $p_h = 0$, to all the events; or even (like *Kb*) forces me to choose some finite subset of them to which I attribute a total probability of at least 99% (leaving 1% for the remainder; and I could have said 99·999% with 0·001% remaining, or something even more extreme). If I do not happen to hold these opinions, and have no reasons for adopting them, then this is no longer a question of coherence; it is a direct interference with my judgement![62]

Moreover, to permit the assignment of zero probabilities to all the events (of a countable partition) is a much less restrictive idea than, as in the finite case, considering them as 'equally probable' ($\mathbf{P}(E_h) = 1/n$, $h = 1, 2,\ldots, n$). The equivalent of this would be to consider the $\mathbf{P}(E_h)$ equal, not in the sense that $\mathbf{P}(E_h) = \mathbf{P}(E_k) = 0$ as real numbers, but in the sense that $\mathbf{P}(E_n)/\mathbf{P}(E_k) = 1$ (as a ratio of zero, or 'infinitely small' probabilities). For the first condition ($p_h = 0$) to hold, much less is required: in terms of ratios, it is sufficient that there do not exist probabilities of maximal order (for example, that give, for each h, $\mathbf{P}(E_{h+1})/\mathbf{P}(E_h) = \infty$), or that, if they do exist, their sum (taking one of them to be unity) is infinite. (It is also sufficient – but this cannot be derived from the ratios alone – that the probability of all the cases be infinite in the given scale; and it may happen that this occurs for the union of cases with probability of smaller order without occurring for those of maximal order.)

Assuming (in line with Axiom 3) that, if accepted, axioms *Zb* or *Kb* should also hold for probabilities conditional on an arbitrary possible event H, they would imply even more restrictive conditions for the probabilities of individual 'possible cases', in order to avoid – pulling out a countable number of them – the possibility of a case of probabilities all zero. There could be at most a countable number with the same order, so that, given a non-denumerable infinity of 'possible cases', we should have a nondenumerable infinity of different 'orders' of probability.[63]

We should also mention that, from time to time, problems which, explicitly or implicitly, run counter to the assumption of countable additivity are also considered by authors who insist on the latter as an axiom. Sometimes the case of 'an integer chosen at random' is regarded as 'meaningful' but 'breaking the rules' (the probability taken to be the limit density; for example, the probability that the integer is a multiple of $k = \lim$ [(number of multiples of k between 1 and n)/n] = $1/k$). At other times (see, for example, Rényi, Chapter 3, 3.18.5), one considers conditional probabilities; for instance, a distribution inside a circle, which is then made larger and larger, so that the probability of each finite region tends to zero. Countable additivity is prescribed for the conditional probability (i.e. within any circle), but it is not made clear that this no longer holds for the limit distribution (which is not explicitly dealt with in its own right, the passage to the limit being merely a device).

This seems to provide further evidence in support of our initial impression that the assumption of countable additivity owes very little to genuine probabilistic

62 It is strange that the very same people who, in general, would encourage one in the finite case to accept a judgement of equal probabilities, on the grounds that a person 'knows nothing', seek to prohibit someone who, on the grounds that he 'knows nothing', would like to make the same judgement in the countably infinite case.

63 One can say even more: those of the same order must possess a convergent sum; the different 'orders', arranged in decreasing order, must form a 'well-ordered' sequence, so that there always exists a 'maximal order'.

considerations; in other words, that it is more a mathematical embellishment than a necessary property of probability. Many other anomalies and peculiarities (one might even say – in a psychological rather than a formal sense – contradictions) strengthen the same impression. The fact that 'equal probabilities' are perfectly acceptable in the finite and continuous (points of an interval) cases, but are not allowed in the countable case, can be explained only by drawing attention to our habit of applying, in particular cases, the most widely used tools (in the countable case we are accustomed to summing series!), rather than adhering to the principle of coherence. The very fact that one treats the finite and uncountable cases differently from the countable case (axioms *Zb* and *Kb*) is sufficient to show that more thought is given to the mathematical structure than to the logical problem – for which the meaningful distinction, if any, would seem to be that between the finite and the infinite (of whatever kind).

Bearing this in mind, the line we have followed here seems to represent, independently of the reasoning we have put forward – which, hopefully, is more persuasive – the most natural way of connecting together attitudes that are, at least in part, inspired by fragmentary and irreconcilable points of view.

Finally, we should mention a concept and a result that have come to be considered as a justification for the systematic use of σ-additivity. The basic idea is the possibility of stretching the interpretation in such a way as to be able to attribute the 'missing' probability in the partition to new fictitious entities in order that everything adds up properly. In some cases, in order to salvage countable additivity, it is even claimed that the new entities are not fictitious, but real. I remember having seen something of this kind in a paper (by Kingman, I think) which involved a probability distribution for discrete processes concentrated in the neighbourhood of a limit case where the process would become continuous.[64] This was taken as an indication of the necessity of including the continuous limit cases among the possible cases, in order to be able to foist upon them the probability missing from the sum.

Using this kind of argument, one could say that if the possible cases were the rationals, and if to each of them is attributed zero probability, then we have demonstrated that the real numbers must also *exist*, and be possible, because they are required as the indispensable support for the probability as a whole (=1).

The more general kinds of considerations we have alluded to are more abstract, although, in the final analysis, they reduce to the same type of argument as is involved in the addition of fictitious entities. In mathematics, this kind of argument or procedure is well known to be fruitful (like, for example, the addition of new points in order to compact a space), but, in our case, events must be events, and not abstractions, if the theory is to preserve a concrete meaning; that is to say 'has some meaning', and this, in the formulations we have mentioned, does not happen.

In fact, it is necessary to have recourse to 'ultrafilters' (and this gives only a theoretical possibility of obtaining the desired result). In any case, this would only hold in a field that has been modified with respect to the original one, and the latter is the only one that we are interested in. I have never seen any application to the study of actual cases (and it seems impossible that it should constitute a simplification, rather than an

64 *Translators' note*. J.F.C. Kingman, 'Additive set functions and the theory of probability', *Proc. Cam. Phil. Soc.*, 63 (1967), pp. 767–775.

unnecessary complication, introduced for the purpose of permitting yet another unnecessary complication, i.e. σ-additivity). It seems to me that the only result has been to encourage people even further to consider just those cases where σ-additivity holds directly, and to ignore the others because they can, in theory, be transformed in such a way as to turn out to enjoy, fictitiously, the property which, in the field one ought to be considering, does not actually hold.

In practice, there are quite different purposes for which consideration of ultrafilters can be useful. In particular, for studying '*agglutinated* probabilities' (i.e. probabilities that cannot be subdivided), which can arise in distributions.[65] Think of the case (which we have already mentioned) of distributions on an ultrafilter: an ultrafilter is a family of events (sets) to which one and only one element of a partition can belong, and we attribute probability 1 to the events belonging to it (and, therefore, probability 0 to the others).

The consideration of *filters* can also be useful if one wishes to analyse further the possibility of dividing up the *missing* probability. In general, given a partition into events E_h, with $\Sigma p_h < 1$, it is sufficient to consider another event B (or a partition, $B_1, B_2, \ldots, B_n$), and to form the partition BE_h (or the partitions $B_1E_h, B_2E_h, \ldots, B_nE_h$) of B (or of B_1, of $B_2, \ldots$, of B_n). The missing probability $1 - \Sigma p_h$ can then be divided up between the filters generated by B and by $\tilde{B}$(or by $B_1, B_2, \ldots, B_n$). Think, in particular, of the mass adherent to a point (in a distribution on the real line).

The probability adherent to the left (or to the right) may be further divided up by considering filters; for example, in the case of the sequence of sets I_n of rationals between $x - 1/n$ and x, one obtains the probability adherent from the left on the rationals, or on the irrationals and so on. Of course, just as knowledge of $F(x)$ is not sufficient to separate possible adherent masses from the concentrated ones, it is even less sufficient for these subdivisions, which have to be established on the basis of other considerations.

18.4. *Concerning what is 'reasonable'.* It would be very difficult to reach any conclusion or to make any constructive progress by attempting to conduct a discussion of this topic with supporters of opposing points of view. Each would first of all attempt to challenge the 'reasonableness' of the assumptions of the others, judging them to be too 'theoretical', lacking any concrete value, and based on the assumption of an absolutely unrealistic degree of precision.

There would be no difficulty for anyone in criticizing the formulations of others, and no doubt anyone making such criticisms against the formulation we have adopted here would find good reasons for so doing. The complications we have considered, however, do not arise, unless one wishes to isolate cases that are in a certain sense 'pathological'. For problems which are 'sensible from a practical point of view', one not only avoids these complications, but also those imposed everywhere *a priori* by the assumption of σ-additivity. The latter are harmful because they go beyond what is required in simple cases and, moreover, are over-restrictive in the complicated cases.

Our criticism of countable additivity on the grounds that it precludes one attributing probability to all events (for example, the extension of Jordan–Peano measure to all the

65 See B. de Finetti, 'La struttura delle distribuzioni in un insieme astratto qualsiasi', *Giorn. Ist. Ital. Attuari*, **XVIII** (1955), pp. 1–14. An English translation of this paper, 'The structure of distributions on abstract spaces', forms Chapter 7 of B. de Finetti, *Probability, Induction and Statistics*, John Wiley & Sons (1972).

sets of the interval [0,1]) is in no way intended as implying that the extension to Lebesgue measure is considered insufficient, or that one actually wishes to go further. On the contrary, it means that we consider it to be usually quite sufficient to confine attention to Jordan–Peano measure, but that, if one wishes to go further, the extension should be neither predetermined, nor ruled out in any way. In other words, Lebesgue measure is just one of the infinite number of extensions to larger families of sets and one should be free to choose any of these if one wishes to make such an extension. Should anyone opt for countable additivity as a matter of preference, there is no objection (just as, in a practical case, it is open to one to choose a distribution possessing a continuous density, rather than a less 'regular' one, without feeling that one is forced to make such a choice by virtue of some law of probability). Any other extension (to all sets) is equally legitimate (in principle: so far as its usefulness is concerned, it is not clear whether the Lebesgue extension should be regarded as useful, once it has been made clear that the consequences one derives from it are not consequences of the initial evaluation by virtue of the 'law' of countable additivity, but rather that they derive from the arbitrary choice of a particular one of the possible extensions of the given evaluation).

On the other hand, a similar criticism can be made at a much more basic and fundamental level. In many approaches, one establishes *a priori* that the probability $\mathbf{P}(E)$ has to be given for all the events E of some given family obeying some given conditions; for example, forming a field (in the above case a Borel field) that is considered fixed once and for all. One can then go on to consider only those problems that belong within that field (considered as a single, closed system, and often referred to as a 'probability space'). It is not then possible to evaluate the probabilities of two events A and B without doing the same for the product AB. But it could well be that sometimes one either has to, or wishes to, proceed (albeit temporarily) without the knowledge or evaluation of $\mathbf{P}(AB)$; or that the family of events initially considered (and 'arbitrary') does not contain all the products. The conclusions hold for all events and random quantities linearly dependent on those events one starts from. Indeed, we could start from random quantities, X, for which the previsions, $\mathbf{P}(X)$, were evaluated: the particular fact of whether all, or just a few, or none, of them are events is in itself irrelevant. The case of events seems simpler and more intuitive only because it is more familiar, as well as being more schematic, and capable of more varied representations (set-theoretic, for example). One may then examine for each problem (with no limitations of any kind) the implications of the evaluations already assumed made, and one can complete them by means of any further evaluations that are required to answer the questions of interest.

There is no need to make use of, or mention, probabilities conditional on events of zero probability (or to compare zero probabilities, as non-Archimedean quantities) except when this might be useful for more careful consideration of delicate situations – where it is otherwise easy to adopt a cursory attitude.

As is illustrated in the cases used as examples, cases which are representative of the general situation, the approach we adopt consists in keeping the treatment at the simplest and most concrete level, adhering to the practical meaning, rejecting assumptions that are not supported by compelling arguments (like that of replacing finite additivity by σ-additivity), rejecting the once-and-for-all fixing of closed structures and, instead, in always open-mindedly allowing the possibility of extending the probability field to be studied, as and when required.

In this sense, we obtain the maximum simplification. There are, however, certain circumstances in which complications do, in fact, arise. This happens when the assumption of σ-additivity leads to simple conclusions (well known from measure theory), which either no longer hold if we abandon the assumption, or require more careful and less 'intuitive' formulations. What should we reply to someone who objects to complications of this kind?

Our reply is that such complications are inherent in the fact that when we speak of probability (or prevision) we are referring to functions **P** which may well be finitely additive (instead of being assumed, *a priori*, by virtue of an 'axiom', to be σ-additive over the field under consideration; or, and it amounts to the same thing, that the term 'event' can only be applied to members of some σ-field, restricted in such a way that **P** is σ-additive over it).

There are, therefore, two possibilities. On the one hand, it may be that we wish to be able to make a statement that holds only under the assumption of σ-additivity; in this case, one can state it as it is, making explicit the assumption that **P** be σ-additive (over some given field, or, often, just over some particular partition). There are no complications, apart from that of stating the assumption, and this has the advantage over the abandoned axiom in that it only requires the assumption of the latter over the minimal field for which it is required. The approach is rather like forcing a person to declare that a property holds for continuous functions, or for functions continuous in a given interval, or at a particular point, when the person is accustomed to stating it as valid for all functions (leaving it to be understood that he is only referring to functions which are continuous everywhere).

Alternatively, it may be that one wishes the statement to be valid without the restrictive hypothesis under which those things that held under σ-additivity continue to be true. In this case, matters become more complicated (unless, for reasons of simplicity, one prefers an inaccurate statement). To put it concisely, there is always the possibility of choice: either stick to the assumption of σ-additivity (no longer considered as an axiom), making it clear that one is doing so, or state things in the form necessary for them to turn out to be true independently of this assumption.

It will suffice to recall various of the cases we have already examined (the reader can, if necessary, refer back to the extensive discussion given in the text); the possibility of masses adherent to a point (instead of concentrated at it) and, in particular, concentrated at infinity (improper distributions); the indeterminacy of $\mathbf{P}(X)$ with respect to distributional knowledge in the case of unbounded distributions; the bogus formulation of the 'strong law of large numbers' and related topics.

In many cases, simple, minor modifications of the kind put forward are sufficient to ensure the validity of a statement, independently of the axiom of countable additivity. Moreover – although this is a matter of taste – they serve, because of their 'finitistic' character, to give a more concrete air to things.

All previous considerations can be regarded as variations on a single fundamental theme: the desirability of basing oneself on axioms that are the weakest (i.e. least restrictive) from a mathematical point of view, because they are from a logical point of view the most securely based (i.e. the least disputable), and which lead to results and statements that are the most secure (i.e. the least disputable).

Let us consider again the introduction of $\hat{\mathbf{P}}(X)$ (Chapter 6, 6.5.7). The mathematical definition is unexceptionable and this, by the standards normally adopted, suffices to render $\hat{\mathbf{P}}(X)$ acceptable *by definition* as the value of $\mathbf{P}(X)$. In contrast, we put forward

arguments whose purpose was to establish the existence of possible reasons for considering this choice as being, in addition, a 'reasonable' one (and this, in a certain sense, is even more important). We were careful, however, not to identify $\mathbf{P}(X)$ with $\hat{\mathbf{P}}(X)$. In fact, $\hat{\mathbf{P}}(X)$ is always just *one* of the possible values for the extension of $\mathbf{P}$ outside the field within which it is uniquely defined by the F (although, in a certain sense, it is the most 'reasonable' extension).

This example, and the discussion arising from it, serves also as an illustration of the kind of attitude that results from the choice of a 'conceptual' approach as opposed to a 'formal' one, in the sense already considered. All the ideas and results are drawn from the *meaning* that lies behind the axioms, and not from the mathematical conventions. In contrast to the tendency towards uniquely determining the extension of certain notions by means of special forms of passage to the limit, our effort consists in not admitting, even inadvertently, any restriction that is not the result of simple finitistic inequalities, and which – one might say – goes beyond the idea of the 'method of exhaustion'.

It is not a question of weighing up, *a priori*, one's preferences for this or that mathematical approach but, on the contrary, of emphasizing the need to choose, in any application, the tools most suited to the nature and meaning of the problem. The nature and meaning must not be distorted or disguised in order to introduce tools of a more or less elegant, sophisticated, or 'fashionable' kind.

18.5. *Countable additivity as continuity*. We return to the topic of countable additivity once again, this time in a different (although equivalent and suggestive) guise. We shall examine some further questions and provide further discussion.

The condition of countable additivity for events (as considered so far) can be expressed in an even more meaningful form as a 'continuity' condition. This is the condition which appears among the axioms given by Kolmogorov (and other authors) in the following form (which we shall call 'axiom' *Kb′*):

> *if* $E_1, E_2, \ldots, E_n, \ldots$ *is a sequence of events, each of which is contained in the preceding one, and whose product is empty* (*i.e. there are no 'elementary outcomes' common to all the* E_n), *then* $\mathbf{P}(E_n) \to 0$ *as* $n \to \infty$.

We can see immediately that this condition is equivalent to countable additivity. Let us write

$$E_1 = (E_1 - E_2) + (E_2 - E_3) + \ldots + (E_{n-1} - E_n) + E_n;$$

all the terms in brackets are events by virtue of the inclusion hypothesis, and the probability of E_1 is the sum of the probabilities of the $(E_h - E_{h+1})$ up to some point, plus the remainder. If the latter tends to zero, as axiom *Kb′* requires, the probability is given by the sum of the series, and countable additivity holds. The argument can be turned around straightforwardly: starting from a sequence $C_1, C_2, \ldots, C_n, \ldots$ of incompatible events, and setting $E_n = C_n + C_{n+1} + \ldots$, we reduce to the preceding case (with $C_n = (E_n - E_{n+1})$); in order that the series of the $\mathbf{P}(C_n)$ converges, the remainder, that is $\mathbf{P}(E_n)$, must tend to zero.

In fact, it is easily seen that *Kb′* leads, in general, to a further property, even more meaningful, and showing more clearly the appropriateness of the term 'continuity'. Note that for any sequence of events, or random quantities, we can consider the *lower limit* and the

upper limit (and also, if these coincide, we can consider the *limit*, their common value), just as in analysis. The fact of whether the values of the sequence are known or not (random) is irrelevant. In particular, in the case of events, $E' = \liminf E_n$ and $E'' = \limsup E_n$ are the events that consist of the fact that a finite number of the E_n are *false* and an infinite number are *true*, respectively (i.e. a finite number take the value 0 and an infinite number take the value 1, respectively). To say that $E_n \to E$ (i.e. that E' and E'' *coincide*), or that the limit E of the sequence E_n exists, is to say that *necessarily*, in the case under consideration, from some N onwards the events E_n are either all true or all false (in other words, it is impossible for infinite sequences of both true and false events to occur).

Well, then: in the case of countable additivity one has

$$\begin{aligned}\mathbf{P}(E') = \mathbf{P}(\liminf E_n) &\leqslant \liminf \mathbf{P}(E_n) \\ &\leqslant \limsup \mathbf{P}(E_n) \leqslant \mathbf{P}(\limsup E_n) = \mathbf{P}(E'');\end{aligned}$$

in particular, if the limit E exists, $\lim \mathbf{P}(E_n) = \mathbf{P}(\lim E_n) = \mathbf{P}(E)$.

It remains to check that the same condition holds more generally when we have a sequence of random quantities X_n rather than events.

The property

$$\mathbf{P}(X_n) \to 0 \quad \text{if } X_n \to 0$$

is valid (under the assumption of countable additivity) *if the random quantities X_n are uniformly bounded*. Suppose, in fact, that, for all n, $|X_n| < K$; then, for any (small) $\varepsilon > 0$, we have

$$|\mathbf{P}(X_n)| \leqslant \mathbf{P}(|X_n|) < \varepsilon + K \,.\, \mathbf{P}(|X_n| > \varepsilon)$$

(because $|X_n| < \varepsilon + K.(|X_n| > \varepsilon) = \varepsilon$ if $|X_n| < \varepsilon$, and $= \varepsilon + K$ otherwise). But if $X_n \to 0$ we also have $(|X_n| > \varepsilon) \to 0$, and this means – recall the above – that we cannot have an infinite number of $|X_n| > \varepsilon$, and hence (assuming Kb') $\mathbf{P}(|X_n| > \varepsilon) \to 0$. It follows that $\lim |\mathbf{P}(X_n)| < \varepsilon$, and, since ε is arbitrary, that

$$\lim \mathbf{P}(X_n) = \lim |\mathbf{P}(X_n)| = 0.$$

Since events are uniformly bounded ($|E_n| \leqslant 1$) the property we have established is equivalent to Kb' (i.e. to countable additivity). If we remove the condition of uniform boundedness, the property does not hold (even if countable additivity holds). Suppose we take a countable partition into events E_n to which we assign nonzero probabilities p_n with sum = 1, and let us consider the sequence of random quantities $X_n = E_n/p_n$ (which are not uniformly bounded). We obtain $\mathbf{P}(X_n) = p_n/p_n = 1 \to 1 \neq 0$, although $X_n \to 0$ (all the X_n but one are, in fact, = 0). By slightly modifying the example, putting $X_n = E_n / p_n^{\alpha}$ for instance, we obtain $\mathbf{P}(X_n) = p_n^{1-\alpha}$, and we see, therefore, that the property $\mathbf{P}(X_n) \to 0$ holds for $\alpha < 1$, whereas, if $\alpha > 1$, we have $\mathbf{P}(X_n) \to \infty$ (although, for the same reason as before, we still have $X_n \to 0$).

The extension of the property to the limit is similar. Assuming Kb, we have

$$\begin{aligned}\mathbf{P}(X') = \mathbf{P}(\liminf X_n) &\leqslant \liminf \mathbf{P}(X_n) \\ &\leqslant \limsup \mathbf{P}(X_n) \leqslant \mathbf{P}(\limsup X_n) = \mathbf{P}(X'')\end{aligned}$$

if the X_n are uniformly bounded. We shall give the proof in the case of the upper limit (the other case is clearly symmetric) and our proof includes the case of events (where a proof was not given).

Putting $X_n'' = \sup X_h$ (for $h \geqslant n$), $X'' = \limsup X_n = \inf X_n''$, we obtain $X_n'' = X_n + (X_n'' - X_n) = X'' + (X_n'' - X'')$, where $(X_n'' - X_n)$ and $(X_n'' - X_n)$ are non-negative. We therefore have, in every case,

$$\mathbf{P}(X_n) \leqslant \mathbf{P}(X'') + \mathbf{P}(X_n'' - X''),$$

and, if $\mathbf{P}(X_n'' - X_n) \to 0$, we shall have lim sup $\mathbf{P}(X_n) \leqslant \mathbf{P}(X'')$. But

$$X_n'' - X_n \to 0,$$

by definition, and, if we assume the X_n to be uniformly bounded (which implies, *a fortiori*, that the $X_n'' - X_n$ are), then, assuming countable additivity, the condition will be satisfied.

In particular, if the sequence of the X_n converges (definitely) to a limit (in general random), $X = \lim X_n$, we can, if we assume countable additivity plus the uniform boundedness of the X_n, state that $\mathbf{P}(X) = \lim \mathbf{P}(X_n)$. The case of a series $\sum X_h$ reduces to the preceding case if we consider the partial sums, $Y_n = \sum X_h$ ($h \leqslant n$). If we call Y' and Y'' the infimum and supremum of the sums, we have

$$\mathbf{P}(Y') \leqslant \inf \sum \mathbf{P}(X_h) \leqslant \sup \sum \mathbf{P}(X_h) \leqslant \mathbf{P}(Y''),$$

and, in particular, we have $\mathbf{P}(Y) = \mathbf{P}(\sum X_h) = \sum \mathbf{P}(X_h)$ if $Y' = Y'' = Y$ (i.e. the series is definitely convergent) under the condition that the remainders are uniformly bounded (and always, of course, with the assumption of countable additivity).[66]

We have said that we do not intend to consider countable additivity as an axiom; for us, it is a property that may appear more or less interesting and which will hold over certain partitions but not over others. Interpreting the condition as one of continuity, we can reformulate this fact in a more meaningful way by saying that it will hold *over certain linear spaces* and not over others.

This approach has the merit of spotlighting the real essence of the problem: the fact that the property of countable or finite additivity, that is of continuity or the absence of continuity, concerns the behaviour of the function **P** over a linear space L. To give a complete account of the behaviour of **P** in terms of continuity involves, therefore, distinguishing which linear spaces L belong to the complex Λ_p of linear spaces over which **P** is continuous, and which do not.

---

66 The convergence of the series $\sum \mathbf{P}(|X_h|)$ (together with the assumption of countable additivity) is a sufficient condition to establish that $\sum X_h$ has probability = 1 of being convergent, and that (putting therefore, arbitrarily, $Y = Y'$, or $Y = Y''$, or $Y = Y' = Y''$, if they coincide, and otherwise $Y = 0$, etc.) one has $\hat{\mathbf{P}}(Y) = \sum \mathbf{P}(X_h)$ (N.B.: $\hat{\mathbf{P}}$ not **P**). Under this hypothesis, in fact, for arbitrary choices of positive ε and λ, there exists an N such that, for any q, we have $\sum \mathbf{P}(|X_h|)(N \leqslant h \leqslant N+q) < \lambda\varepsilon$, i.e. $\mathbf{P}\{\sum |X_h| (N \leqslant h \leqslant N+q)\} < \lambda\varepsilon$, and, *a fortiori* (denoting the preceding summation by $\{\Sigma_{N,q}\}$ for short), $\mathbf{P}(\Sigma_{N,q} > \lambda) < \varepsilon$ (because, if X is certainly positive, $(X > \lambda) \leqslant X/\lambda$). If the axiom of continuity holds, as we have assumed, the limit-event $(\Sigma_N > \lambda) = \lim(\Sigma_{N,q} > \lambda)$ (as $q \to \infty$) also has probability $\leqslant \varepsilon$, and, *a fortiori*, it follows that the fact that the series diverges (in which case the remainder Σ_N will be ∞) has probability $\leqslant \varepsilon$, and hence (since ε is arbitrary) *zero*.

More precisely, in line with what we have said previously, and with the condition of coherence, we shall say that **P** is *coherent and continuous* on L if no random quantity X of the form

$$X = k_1\left(X_1 - \mathbf{P}\left(X_1\right)\right) + k_2\left(X_2 - \mathbf{P}\left(X_2\right)\right) + \ldots + k_n\left(X_n - \mathbf{P}\left(X_n\right)\right) + \ldots$$

turns out to be uniformly positive (where the k_h are any real numbers and the X_h belong to $\mathscr{L}$), not only for sums involving a finite number of terms (as is required for coherence) but also for series (convergent, and with uniformly bounded remainders).

It is clear that if $\mathscr{L}$ belongs to Λ_p then so does every linear space contained in it, and so does the closure, $\bar{\mathscr{L}}$, formed by all the random quantities that can be obtained from $\mathscr{L}$ by means of the passage to the limit in the sense given:

$$\left(X_n \to X, |X_n - X| < K\right).$$

If $\mathscr{L}_1$ and $\mathscr{L}_2$ belong, then so does $\mathscr{L}_1 + \mathscr{L}_2$ (the linear space of sums $X_1 + X_2$, $X_1 \in \mathscr{L}_1$ and $X_2 \in \mathscr{L}_2$): this holds for any finite number of spaces L_h *but not for an infinite number*.[67] This indicates that the most 'natural' hypothesis is not true; a hypothesis which corresponds most closely to the standard point of view because it leads to a distinction between those events and random quantities which belong to a certain system of 'probabilizable' entities, and those which do not. This is the hypothesis that the complex Λ_p consists of all and only those linear spaces that belong to some given linear space L*, which, in this case, would have acquired the meaning of 'total field of continuity'.

19 Questions Concerning Qualitative Formulations

19.1. There are many senses in which the words qualitative probability have been used, some of them very different from each other. To attempt to list them and classify them would be both tedious and pointless, but something must be said in order to point out the necessity of not confusing things that do differ, and of not being put off by apparent absurdities. Among the latter, for example, we include the fact that one might expect to encounter rather vague considerations, but can, in fact, find oneself forced into hair-splitting detail, obliging one to apply, in all cases, the methods of comparison introduced for zero probabilities.

Our day-to-day judgements are on the whole rather vague and we usually limit ourselves to just a few verbal gradations (quite probable, or very, very much, not much, very little,…), or to percentage approximations (50%, 75%, 90%, 99%,…). In comparisons between two events, the probabilities will be said to be 'roughly equal' if the dominance of one over the other does not appear to be obvious. At this level, however, there is not even the possibility of arguing in mathematical terms.

---

67 Consider a countable partition of events E_{hk} ($h, k = 1, 2,\ldots, n,\ldots$). Let us denote by $E_h = \sum_k E_{hk}$ the sum of events whose first subscript is h, and suppose we attribute the values $p_{hk} = \mathbf{P}(E_{hk})$ and $p_h = \mathbf{P}(E_h)$ in such a way that $\sum_k p_{hk} = p_h$ (for each h), but $\sum_h p_h < 1$ (i.e. $\sum_{hk} p_{hk} < 1$). On the linear spaces $\mathscr{L}_h$ defined by the E_{hk} and E_h, **P** is continuous, and hence is also continuous on every linear space $\mathscr{L}$ determined by a finite number of L_h. This no longer holds, however, if we consider the space $\mathscr{L}$ determined by the whole infinite collection of L_h.

Sometimes, one thinks of vagueness in the sense of 'indeterminacy' (for example, between precise numerical bounds); we have already referred to this, and we shall return to it later. At other times, one is willing to compare (let us assume exact comparison, in order not to get lost in too many subcases) the probabilities of events but without using numerical probabilities. A physician might have a quite precise opinion, in a comparative sense, concerning the probabilities that a small number of patients will overcome their present disease, but without knowing what to do if he were required to compare them with the probability of obtaining something other than a '6' on the role of a die (or, more explicitly, were he required to state whether they were more or less than $\frac{5}{6}$; i.e. 83·3%). Sometimes, this inability to compare them with numbers is attributed to innate peculiarities of the events in question (rather than to contingent reasons, such as lack of practice; see Borel's review of Keynes' treatise[68]), or to the fact of not having at one's disposal (or not wishing to have) devices such as dice, urns and so on. In this case, if comparability is assumed exact, as in the comparison of intervals where one is led to say that a closed interval (i.e. end-points included) is greater than one of equal length but open (i.e. end-points excluded), it is clear that a non-Archimedean scale results (and this is the absurdity we referred to at the beginning).

Other considerations arise when indeterminacy has a precise meaning; when, on the basis of some data, one can establish only that a probability p belongs to an interval $p' \leqslant p \leqslant p''$. That we are not dealing with an essential indeterminacy is a point that we have stressed. Nevertheless, there is something to be said here and a few points will have to be made in connection with the discussion (in Sections 5–7) relating to the verifiability of events and measurement of quantities.

19.2. *Axiomatic formulations in qualitative form.* In all the methods of approach we have so far looked at, we have introduced, straightaway, numerical values for probabilities under the intuitive guise of prices, and as parameters required for optimal decision making. In so doing, we referred (albeit indirectly) to percentages of white balls, or to number of successes, and so on. This is certainly the most direct way of learning how to express one's own opinions and how to formulate the mathematical conditions which they must satisfy (and in terms of which they can be manipulated in a probabilistic argument).

There are occasions, on the other hand, when it seems preferable to start from a purely ordinal relation – that is a qualitative one – which either replaces the quantitative notion (should one consider it to be meaningless, or, anyway, if one simply wishes to avoid it), or is used as a first step towards its definition. For example, given two commodities (or two economic alternatives) A and B, one can ask which is preferable (or whether they are equally preferable) before defining utility (or perhaps even rejecting the very idea of measurable utility); and the same can be said for temperature, the pitch of a note, the length of intervals and so on.

One could proceed in a similar manner for probabilities, too. In fact (if one accepts the subjective point of view), one can apply precisely the same notion of preference as we mentioned for utility. Instead of two commodities A and B, one compares one and the same gain (let us say 1 lira) conditional on the occurrence of event A, or event B.

---

68 The article is reprinted in Borel's *Traité* (as note 2 in issue III of Vol. IV); an English translation is given in H.E. Kyburg and H.E. Smokler, *Studies in Subjective Probability*, John Wiley & Sons, Inc., New York (1964).

Our preference (apart from reservations concerning 'distorting factors'; see Section 13 above) will be for the event judged more probable (or, if the two events are judged to have the same probability, we will be indifferent).

This approach has been studied, and, provided one does not insist on splitting hairs, leads quickly and naturally to the usual conclusions (although in a form less directly applicable to the general case). The properties one needs to take as axioms are simple and intuitive (the standard order properties, plus the qualitative equivalent of additivity): given that E' and E'' are incompatible with E, then $E \vee E'$ is more or less probable than $E \vee E''$, or equally probable, according to whether E' is more or less probable than E'', or equally probable; in other words, logical sums preserve order.[69] All the same, the 'qualitative' comparison inevitably turns out to be far too precise (indeed, far too sophisticated), from a theoretical point of view, for what is required for the quantitative (numerical) evaluation; in any case, it is not conveniently translatable into such an evaluation (unless one considers the possibility of constructing special scales of comparison).

The complication derives from the fact that, in a qualitative sense, a possible event (no matter what probability p one attributes to it, even $p = 0$) is obviously 'more probable' than an impossible event. Similarly, by adding to an event E a possible event A, incompatible with it, even if of zero probability, one obtains an event $E + A$, which is 'more probable' than E. It follows that, having other events E' with (numerical) probabilities equal to that of E, $\mathbf{P}(E') = \mathbf{P}(E)$, the qualitative comparison would have to establish for each one whether it had the same probability as E, or $E + A$, or greater than the first and less than the second, or greater than both, or less than both. Even worse; consider an arbitrary sequence of events $A_1, A_2, \ldots, A_h, \ldots$, all of zero probability, mutually incompatible, and incompatible with E, and another sequence of events $B_1, B_2, \ldots, B_h, \ldots$, all of zero probability, mutually incompatible, and contained in E: setting

$$E_0 = E, \quad E_h = E + A_1 + A_2 + \ldots + A_h,$$
$$E_{-h} = E - B_1 - B_2 - \ldots - B_h, \quad (h > 0),$$

one obtains an increasing ($E_h \subset E_k$ for $h < k$) and doubly unbounded sequence of events E_h ($h = 0, \pm 1, \pm 2, \ldots, \pm n, \ldots$), all with probability $\mathbf{P}(E_h) = \mathbf{P}(E)$. Any comparison of an E' (also having probability $\mathbf{P}(E') = \mathbf{P}(E)$) with the E_h should make precise which (if any) of the E_h have the same probability as E'; or, otherwise, in which of the intervals E_h, E_{h+1}, it finds itself; or if it precedes, or follows, all the E_h, for h between $\pm\infty$.

The necessity, now explained, of this much more refined comparison, has led us to use the phrase 'having the *same* probability' for two events which, in the ordering, belong to the same 'equivalence class', rather than 'equally probable', which we use when we refer to the equality of their numerical probabilities.

The situation is that which would present itself (in a less serious way) in a comparison between intervals, if intervals of equal length were to be called 'equally long' only if they both contained either 0, 1 or 2 of their end-points (otherwise, the one containing more end-points would be called 'longer'). One arrives at a closer analogy by extending the example to sets which are unions of a finite number of intervals, and (the sum of the

69 See B. de Finetti, 'Sul significato soggettivo della probabilità', *Fundamenta Mathematicae*, 17, Warsaw (1931); an improvement in the argument given in the notes of my course on the Calculus of Probability, University of Padua, 1937–1938, was made by Professor A. Gennaro who succeeded me in presenting the course.

lengths being equal) calling 'longer' the set for which the difference between the number of closed and open components is greatest (the intervals containing only one end-point are not counted; any isolated points are counted as closed intervals; these conventions are necessary if we are to have additivity, as in the case of probability).

Using such partitions into intervals as an image for our probabilistic partitions, one sees, for example, that, if the certain event is thought of as represented by a closed interval of length 1, it is impossible to divide it into two intervals (or, more generally, into n) that have the same probability (there are $n + 1$ end-points, one too many). This difficulty cannot be overcome by changing partitions into sums of intervals: it is always a question of dividing up the length 1 (into intervals with one end-point), plus one point. Shifting an end-point from one of the intervals to another one creates a disparity between them, but, in total, there always remains one end-point too many.

Conversely, if one does not consider that one has (included in the field of events to be compared) events that are suitable for furnishing a scale of comparison (for example, drawing balls numbered 1 to n, and judged to have the same probabilities, from an urn, with n arbitrarily large[70]), then the inequalities arising can provide totally inadequate information about the numerical values of the probabilities. Given, for example, a partition into three (incompatible) events A, B, C (assumed in order of decreasing probability), and that the only remaining comparison open to us is between A and $B + C$, this will tell us whether the probability of A is greater than or less than $\frac{1}{2}$. In the former case, we know only that $\mathbf{P}(A)$ lies between $\frac{1}{2}$ and 1, $\mathbf{P}(B)$ between 0 and $\frac{1}{2}$, $\mathbf{P}(C)$ between 0 and $\frac{1}{4}$; in the latter case, we know that $\mathbf{P}(A)$ lies between $\frac{1}{3}$ and $\frac{1}{2}$, $\mathbf{P}(B)$ between $\frac{1}{4}$ and $\frac{1}{2}$, $\mathbf{P}(C)$ between 0 and $\frac{1}{3}$ (Figure A.2). And it cannot be said that things necessarily improve if we consider more than three events. If, for example, the most probable of them is more probable than the union of the others, one can only say that its probability lies between $\frac{1}{2}$ and 1; the others, therefore, in aggregate, can have probability close to $\frac{1}{2}$, or arbitrarily close to zero, or even zero.[71]

We could avoid complications of this kind by assigning to the comparison 'A is more probable than B', a meaning equivalent to $\mathbf{P}(A) > \mathbf{P}(B)$, and, in particular, calling a possible event of zero probability 'equal in probability to the impossible event'. In order to do this, it would be necessary to introduce the Archimedean property; in other words, to characterize those events 'more probable than the impossible one' as those with a positive numerical probability by means of a condition like the following: 'there exists a finite N such that, in every partition into N events, at least one is less probable than the given event' (whose probability is then $\geqslant 1/N$). But why resort to this distortion of an arithmetic condition instead of proceeding directly, given that one's desired goal is, in fact, the arithmetic notion?

---

70 The inconvenience of having to postulate the existence of partitions into events having the same probabilities has been overcome in L.J. Savage, *The Foundations of Statistics* (Chapter 3, 3: 'Quantitative personal probability') by means of a weaker assumption: for any N, one can construct a partition into N parts such that no union of n parts is more probable than one of $n + 1$ parts (for any $n < N$).

71 In the case of $n > 3$ parts (and in the absence of a 'scale of comparison') it becomes complicated to even establish the compatibility of a system of inequalities (between sums of events of the partition). For $n = 4$, there is a simple sufficient condition (as I showed in my paper 'La logica del plausibile secondo la concezione di Pólya', *Atti. Riun. S.I.P.S.* 1949 (1951)). Contrary to what I had supposed, however, this does not hold for $n > 4$, as was shown by C.H. Kraft, J. Pratt and A. Seidenberg, 'Intuitive probabilities on finite sets', *Annals of Mathematical Statistics*, **XXX** (1959).

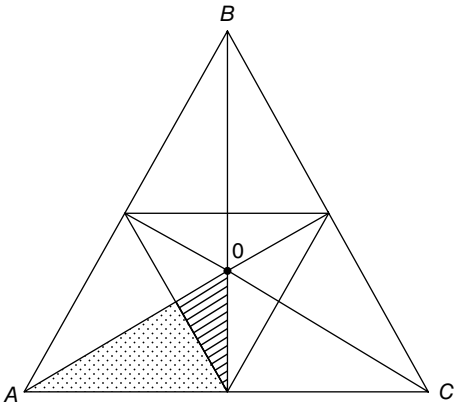


Figure A.2 Areas distinguished by means of the comparisons of probabilities for the events of a partition and their sums. The above refers to the case $n = 3$.

Dotted area: $A > B > C$, $A > B + C$.
Shaded area: $A > B > C$, $A < B + C$.

For $n = 4$ (tetrahedron), and $n > 4$ (simplex in higher-dimensional space), the mode of subdivision is similar.

On the other hand, the 'sophisticated', non-Archimedean, criterion corresponds exactly to the purely logical meaning that one would like to give to the comparison of probabilities in the particular case in which it relates to an objective condition, that of implication. If it is true (indeed, if it is certain; i.e. if we know) that $A \subset B$ (an objective condition), then coherence obliges us to evaluate $\mathbf{P}(A) \leqslant \mathbf{P}(B)$ (subjective evaluation). One can say that A is 'less probable in an objective sense' (or 'less possible') than B if and only if $A \subset B$, because if A occurs then B certainly occurs, and, moreover, it is possible that B occurs without A occurring. Viewed in this light, the fact that the event $B - A$ will be assigned zero probability under some evaluations, and nonzero probabilities under others, is irrelevant: the important thing is that $B - A$ is possible. To give a geometrical analogy in this case, also, one could say that a comparison between two sets in the absence of a notion of a metric can only lead one to assert that A is smaller than B when the former is properly contained within the latter. In this case, and only in this case, will it be true that $m(A) \leqslant m(B)$ for *whatever* measure m might be introduced (the exact form being $<$ or $=$, according to whether the measure in question attributes to $B - A$ a positive or zero value).

Given disparate events E' and E'', it seems too much to hope that by comparing their probabilities one can decide if $\mathbf{P}(E') = \mathbf{P}(E'')$ exactly, rather than whether there is a difference of about 10^{-6}, or 10^{-1000}, and so on. Should one wish to square the reasonableness of the procedure with the logical scruples mentioned above (i.e. the seemingly obvious fact that, for the same price, one prefers to have an extra possibility of winning, even though the extra possibility has probability zero), one could perhaps consider an intermediate order relation. One might define (for example):

$A < B$: 'A less probable than B'	if $\mathbf{P}(A) < \mathbf{P}(B)$, or if $\mathbf{P}(A) = \mathbf{P}(B)$ and $A \subset B$;
$A \sim B$: 'A not comparable with B'	if neither $A < B$ nor $B < A$; that is if $\mathbf{P}(A) = \mathbf{P}(B)$ and we have neither $A \subset B$ nor $B \subset A$.

Instead of following this rather abstract comparison of the various possibilities, it is more useful to ask oneself whether the exigencies of the problem themselves indicate

the appropriate form, which could then profitably be adopted. This approach leads to something quite similar to the above, but, in the case where $\mathbf{P}(A) = \mathbf{P}(B)$, comparability will be given by a condition less restrictive than demanding that either $A \subset B$ or $B \subset A$ (i.e. either $A - AB = 0$, $\mathbf{P}(B - AB) = 0$, or vice versa). Specifically, the condition consists of $\mathbf{P}(A - AB) = \mathbf{P}(B - AB) = 0$, together with the comparability of these two zero probabilities, along the lines suggested in Chapter 4 and developed in Section 18.2 of this Appendix.

It is a question of comparing the conditional probabilities

$$\mathbf{P}\big(A - AB \,|\, A + B - AB\big) \quad \text{and} \quad \mathbf{P}\big(B - AB \,|\, A + B - AB\big)$$

(whose sum = 1), saying that A is more or less probable than B according to whether the first expression is greater than the second or is less. Note that in this way one can deal with every case within the one formulation; it does not matter whether $\mathbf{P}(A)$ and $\mathbf{P}(B)$ are different, or are equal, or whether, if they are zero, one has to proceed to the comparison of residuals. By definition, we have that A and B are not comparable if $\mathbf{P}(A) = \mathbf{P}(B)$ and the residuals $A - AB$ and $B - AB$ have equal probabilities (which is automatic if they are not zero).

Note also that it would be wrong to claim that events which are not comparable have the same probabilities. If A is more probable than B on account of the zero probability of $A - AB$ being greater than that of $B - AB$, and if C is an event such that $\mathbf{P}(C) = \mathbf{P}(A) = \mathbf{P}(B)$ but $\mathbf{P}(AC) < \mathbf{P}(C)$, then both A and B turn out not to be comparable with C; but to say 'having the same probability as C' would imply that they had the same probability as each other, and this is false by hypothesis.[72]

19.3. *Do 'imprecise probabilities' exist*? The question as it stands is rather ill-defined, and we must first of all make precise what we mean. In actual fact, there is no doubt that quantities can neither be measured, nor thought of as really defined, with the absolute precision demanded by mathematical abstraction (can we say whether the number in question is algebraic or transcendental? Or are we capable of giving millions of significant figures, or even a few dozen?). A subjective evaluation, like that involved in expressing a probability, attracts this criticism to an even greater degree (but this is no reason for regarding the problem differently in this case, as somehow being more essentially rooted in the concepts involved). The same is true for 'objective probability': the person putting his faith in 'objective probabilities' is in precisely the same situation, except insofar as he is restricting himself to cases in which everyone (he himself, or even a subjectivist) has at his disposal criteria and information which make the judgement easier.

In this sense, it should be sufficient to say that all probabilities, like all quantities, are in practice imprecise, and that in every problem involving probability one should provide, just as one does for other measurements, evaluations whose precision is adequate in relation to the importance of the consequences that may follow. In any case, one should take into account that there is always this margin of error (for instance, it might be worth repeating the calculations with several slightly different values).

---

72 The only case in which the two events E' and E'' could be said 'to have the same probability' is that in which they consisted, respectively, of m' out of n', and m'' out of n'', events of two partitions into events having the same probabilities, where $m'/n' = m''/n''$. This remark should not be taken to mean that we wish these futile and absurd complications to be taken seriously, but, on the contrary, that we wish to remove them, without ignoring, however, the issue of what can or cannot be expressed in a correct way.

The question posed originally, however, really concerns a different issue, one which has been raised by several authors (each of whom, it seems to me, imparts a different shade of meaning to the problem). It concerns the possibility of cases in which one is not able to speak of a single value p for a given probability, but rather of two values, p' and p'', which bound an area of indeterminacy, $p' \leqslant p \leqslant p''$, possessing some essential significance.

The idea can be traced back to Keynes (see the remark in the last section concerning Borel's review), and was later taken up by B.O. Koopman and I.J. Good, developed considerably by C.A.B. Smith,[73] and more recently by other authors, like Ellsberg and Dempster.

Several different situations may lead one to express oneself in terms of an imprecise evaluation.

An example of this occurs when one wishes to distinguish various hypotheses, and attributes different probabilities $\mathbf{P}(E|H_i)$ to an E, depending on the various hypotheses H_i; if one then ignores the hypotheses, one can only conclude that the probability lies between the maximum and the minimum. We have already dealt with this case in Chapter 4, Section 4.8, especially in 4.8.3 and 4.8.5. The probability is what it is on the basis of the information that one has. It is clear that with additional information the probability could take on all conceivable values, finally reaching, and then remaining at, either 1 or 0, when it is finally known whether it is true or false. If we are dealing with hypotheses H_i about which we expect soon to have some information, then it may be reasonable to wait until this information is available, rather than making a provisional evaluation by taking a weighted average with respect to the probabilities which, in the meantime, are attributed to the H_i. It would be naïve, however, to assert that $\mathbf{P}(E)$ will take on a value lying somewhere between the $\mathbf{P}(E|H_i)$. There is an infinite number of partitions in hypotheses and the information which comes along might be anything at all (for instance, it may confirm that out of the H_i, H'_j, H''_l those with $i = 3$, $j = 1$, $l = 7$ are true); the $\mathbf{P}(E \mid H_3 H'_1 H''_7)$ could vary anywhere between 0 and 1, even though the $\mathbf{P}(E|H_i)$ are all very close to one another (and even if they are equal; this case is only without interest insofar as the question would then, of course, never have been raised).

A second example occurs when one has not given sufficient thought to the matter, and hence possesses only a vague idea of the evaluation one wishes to make. A special case of this occurs when one has expressed the evaluation in terms of a formula (e.g. $p = \mathrm{e}^{-a}\mathrm{a}^{\mathrm{n}}/n!$, with $a = 5813$ and $n = 12$), but, having not yet carried out the numerical calculations, one has only a rough idea of the order of magnitude. In the final analysis, however, nothing has changed. One either carries out the calculations or one is obliged to take as the probability the prevision of the result according to one's own, more or less haphazard, crude estimation: there is no better solution.

The idea of translating the imprecision into bounds, $p' \leqslant p \leqslant p''$, even in the weaker sense proposed by Good (who regards p' and p'' not as absurd, rigid bounds, capable of

73 These questions are examined in detail in Sections 26 and 27 of the paper by B. de Finetti and L.J. Savage that we have frequently referred to; particular reference is made to the (then very recent) paper of C.A.B. Smith, and to the interesting discussion to which it gave rise at a Royal Statistical Society meeting, with contributions from Barnard, Cox, Lindley, Finney, Armitage, Pike, Kerridge, Bartlett and (in the form of a written contribution) Anscombe. The reader will find there many other points which we have not found room for here.

'making the imprecision precise', but merely as indications of maxima), is inadequate if one wishes to take it to the limit in the sense in which it serves to give an idea of the imprecision with which every quantity is known or can be considered. One should think of the imprecision in the choice of the function **P** (extended, for example, to some neighbourhood P* of a given $\mathbf{P}_0$ in the space P).The imprecision for individual events and random quantities would, as a consequence, be determined not as isolated features, but with the certain or uncertain connections deriving from logical or probabilistic relations.

19.4. *What one can do in practice.* A similar kind of discussion can be given concerning what one can actually do in practice; the main purpose of this, however, is to make clear that the issues involved here are rather different and do not give rise to any difficulties or anxieties.

The (theoretical) possibility of attributing a *precise probability* to all the events appears to be an indispensable requirement if one considers probability as a notion which applies to events per se, independently (see Section 3) of the existence and nature of any properties (e.g. topological) of the fields to which the definition of certain events can be referred. On the other hand, this does not imply that these probabilities are determined, as unique extensions of those conferring to a subfield (extensions which may or may not provide a sharpening of bounds; sometimes, as a special case, they may provide a unique value), nor does it imply that we are obliged to complete the evaluation, nor even to worry about it. Indeed, one can even call a halt well before this (earlier than usual) if there is no real interest in proceeding further: this is even more the case if the hypotheses upon which one would base oneself in proceeding further appear rather artificial and devoid of any realistic meaning.

If, in the context of the above (Section 19.3), we stop thinking in terms of a certain subset $\mathscr{P}^*$ of previsions $\mathbf{P} \in \mathscr{P}^*$, among which we are uncertain as to which one to choose, and we consider instead that we are dealing with the set of all the $\mathscr{P}$ which extend a given **P** in $\mathscr{L}$, then the bounds of indeterminacy mentioned above (at the end of Section 15) would follow. But it is not a question of imprecision. The fact that, in the case we are considering, it is only possible to say of a $\mathbf{P}(E)$ that it lies between $\mathbf{P}^-(E)$ and $\mathbf{P}^+(E)$, does not imply that certain events, like E, have an indeterminate probability: it merely implies that the probability is not uniquely defined by the initial data that one has considered. As an analogy, it would, in the same way, be nonsense to say that 'a rectangle having a perimeter of 12 m has an indeterminate diagonal'; it is 'determinate' in the sense that it is what it is, but one has to measure it, or measure a side, or something else, in order to obtain sufficient information to be able to 'determine' it (in the sense of obtaining, by means of a calculation, its well-determined value, notwithstanding the fact that knowledge of the perimeter alone is not sufficient).

There is a context in which one might refer to indeterminacy, but only in a precise, technical sense. This would arise if a certain individual were familiar with the evaluations that another individual had made within $\mathscr{L}$, and wished to establish what *the latter* should do outside that ambit in order to remain coherent. This makes it clear, however, that the indeterminacy is not a property of the events, but rather that it lies in the fact that an outsider cannot remove it, since he cannot replace the individual who is interested in the, as yet unprejudiced, evaluation.

Another aspect of the problem links up with the discussion (Sections 5–11) concerning the 'verifiability' of events (or the 'realizability' of measurements).

If, for a given problem, in a given situation, an event is not, in practice, verifiable, then any discussion about its probability is mere idle talk. For this reason, leaving aside the question of whether or not one accepts the necessity of admitting countable additivity as an axiom, it seems that if one is discussing the probability that $X \in I$, where I is non-measurable in the Jordan–Peano sense, and there is an interval in which both I and its complement $\tilde{I}$ are everywhere dense, then no measurement – not even with the assumption of 'unbounded precision' (Section 7) – can decide, on the basis of an observation x lying in that interval, whether or not the exact value lies in the interval. The same difficulty remains even under weaker assumptions: for example, if there exist points of both I and $\tilde{I}$ whose distance from x is smaller than the margin of measurement error.

In other fields, too, for realistic applications it seems much more useful to use methods which avoid too rigorous an assumption of precision. For example, when one speaks of 'convergence' it seems preferable, when considering results to have asymptotic validity, to confine attention to those which are valid for a large, but finite, number of cases (without taking seriously the notion of considering an infinite number of them).

20 Conclusions

Can one, having now come to the end, draw some conclusions?

I have in mind, of course, the critical questions that we have examined in this Appendix (some of which we anticipated in the text, to the extent that the topic in question required). In other words, I am referring to questions of a predominantly technical nature – if the word technical is adequate to characterize the difference between these matters and those concerning the meaning and formulation of the entire (subjective and Bayesian) theory; matters which occupied us throughout the text (Chapters 1–12).

Some kind of summary is required, if only to avoid the possibility of the reader being left with a feeling of confusion or bewilderment. The latter is a distinct possibility, because of the apparent contrast between our tendency on the one hand to simplify things, refusing to go beyond the level of practical applications, and towards, on the other hand, throwing ourselves headlong into hair-splitting and complicated analyses (which are not only far removed from any foreseeable application, but even strain the limits of good sense).

Why then, someone will surely ask, not be content with the 'happy medium' provided by the standard approach? This consists in proceeding to the point where countable additivity makes everything work beautifully, and then stopping when the miracle ceases.

Because – I would answer – so far as I am concerned it is by no means a 'happy medium', but rather a case of 'two wrongs not making a right'. In my opinion, anything in the formulation that proceeds beyond what the Jordan–Peano–Riemann machinery provides is irrelevant for practical purposes, and unjustifiable on theoretical and conceptual grounds.

The two kinds of discussion to which we referred above, although apparently in contrast to one another, are intended, converging from opposite directions, to demonstrate one and the same point: *one can do without complications* (and this is perhaps the wisest course of action), but, should one decide to embark upon them, *one must do so wholeheartedly, in a constructive manner, even though this may prove troublesome*.

I may be wrong. My criticisms will not have been in vain, however, if in order to refute them someone comes forward and explains and justifies, in a sensible and meaningful way, those things which, up until now, have merely been 'Adhockeries for mathematical convenience'.

Index

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.