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

$$\mathbf{P}(E) = \mathbf{P}^0(EA)/\mathbf{P}^0(A) = \mathbf{P}^0(E)\mathbf{P}^0(A|E)/\mathbf{P}^0(A). \tag{11.1}$$

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, $E_1, E_2, \ldots, E_{50}$, and that the results are the following:

$$1111001111 \qquad 0111000111 \qquad 1100001110 \qquad 0000111011 \qquad 1000000010,$$

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 **P*** for which, given any arbitrary $\varepsilon > 0$, there exist **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 **P** to which it leads as a result of (arbitrary) observations A (the coefficients varying for different experiences, A). In fact, we have

$$\begin{aligned}\mathbf{P}(E) = \frac{\mathbf{P}^0(EA)}{\mathbf{P}^0(A)} &= \frac{c_1^0\mathbf{P}_1(A)\mathbf{P}_1(E) + \ldots + c_m^0\mathbf{P}_m(A)\mathbf{P}_m(E)}{c_1^0\mathbf{P}_1(A) + \ldots + c_m^0\mathbf{P}_m(A)} \\ &= c_1\mathbf{P}_1(E) + \ldots + c_m\mathbf{P}_m(E),\end{aligned} \tag{11.3}$$

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

$$\mathbf{P}\left(\frac{Y'}{n'} - \frac{Y''}{n''}\right)^2 = \frac{n' + n''}{n'n''}\left[\left(m^2 + \sigma^2\right) - \left(m^2 + r\sigma^2\right)\right] = \left(\frac{1}{n'} + \frac{1}{n''}\right)\sigma^2(1 - r). \tag{11.4}$$

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) → 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):

$$K\int_0^1 \xi.\xi^r(1-\xi)^{n-r}\,\mathrm{d}\xi=(r+1)/(n+2)$$
$$\left(K=\left[\int_0^1 \xi^r(1-\xi)^{n-r}\,\mathrm{d}\xi\right]^{-1}\right). \tag{11.10}$$

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

$$p_r^{(n)} = \left[\frac{\omega_{r+1}^{(n+1)}}{\binom{n+1}{r+1}}\right] \Bigg/ \left[\frac{\omega_r^{(n)}}{\binom{n}{r}}\right] = \frac{r+1}{n+2} \Bigg/ \left[1+\left(1-\frac{r+1}{n+2}\right)\left(\frac{\omega_r^{(n+1)}}{\omega_{r+1}^{(n+1)}}-1\right)\right] \tag{11.11}$$

(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:

$$v(\xi)=v_1(\xi)v_2(\xi)\ldots v_n(\xi).$$

Let $\bar{\xi}$ denote the 'maximum likelihood' point: that is the point at which $v(\xi)$ has an absolute maximum (which we shall assume to be unique; and we further assume that $v(\xi)$ is much less than $v(\bar{\xi})$, 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 $v_i(\xi)$ by the linear approximation

$$v_i(\bar{\xi})+v_i'(\bar{\xi})(\xi-\bar{\xi})=K\left[1+a_i(\xi-\bar{\xi})\right]$$

(with $\sum_i a_i = 0$, in order that $v'(\bar{\xi})=0$), 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 $v_i(\xi)$ the factor $f(\xi)$ = prior density, the product becomes

[20] In order to obtain the exact value of $v''(\bar{\xi})$, one must take into account $v''(\xi)$, by writing

$$v_i(\xi)\simeq K\left[1+a_i(\xi-\bar{\xi})+\frac{1}{2}b_i(\xi-\bar{\xi})^2\right].$$

This then gives the approximation

$$v(\xi)=K\left[1-\frac{1}{2}(\xi-\bar{\xi})^2\sum_i\left(a_i^2-b_i\right)\right].$$

(Note that $\sum_{i\neq j}a_ia_j=\sum_i a_i\sum_j a_j-\sum a_i^2$).

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.