12

Mathematical Statistics

12.1 The Scope and Limits of the Treatment

12.1.1. A brief account of mathematical statistics within the confines of the final chapter of this book must necessarily offer a limited perspective. Nevertheless, its inclusion serves a very definite purpose.

In Chapter 11, we have already encountered certain of the problems that fall within the purview of the subject matter of mathematical statistics. Specifically, we examined applications of inductive reasoning based on statistical data; that is, data involving a number of observations (possibly a large number) that are, in a certain sense, similar to one another. We have also explained the Bayesian approach to such problems (an approach which constitutes an integral part of the subjectivistic conception), noting that a unified coherent structure cannot be maintained if this is abandoned in favour of other approaches involving a variety of more or less empirical 'ad hoc' methods.

In this chapter, we shall attempt to give a more explicit account of the problems with which mathematical statistics is concerned, and of the implications, both for these problems and more generally, which flow from the adoption of one or other of the competing points of view.

12.1.2. In addition to the strictly probabilistic aspects, with which the previous considerations are concerned, we shall have occasion to examine other topics relating to decision theory. There are two basic reasons for this and they correspond to two different questions that can be posed within the same framework.

The first of these concerns applications where the entire enterprise is more or less explicitly geared to arriving at a decision. Obvious examples of this are batch testing, or quality control, performed on a sample of a certain product in order to decide what to do with the rest of the stock (whether or not to reject it), or how to produce it in an optimal fashion (whether or not process parameters need adjusting) and so on. Although one does not always have such immediate actions in mind, it could be said that there is always some practical purpose for which one is somehow seeking guidance.

The second reason may or may not be relevant, depending on the particular application: more precisely, it depends on whether or not one is able to experiment. If this is possible – that is if one has certain choices regarding the way in which observations are to be obtained – then there is an additional dimension to the problem, because this choice is itself a decision and must be made in the most appropriate way. What is

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

appropriate will in this case, of course, depend on the final decision, itself dependent on the outcome of the experiment. More precisely, the whole question of what is appropriate needs to be set in the context of the theory of compound decisions; a framework including both decisions concerning the experiment to be performed and the final decision to be taken after the results of the experiment are known.

12.1.3. General issues will be dealt with in a somewhat summary fashion in what follows, and most of the discussion will centre around specific examples. In this way, we hope to provide a straightforward account that will make best use of the limited space available to use. The examples will, in fact, involve some of the most important and commonly occurring cases, and so, in a sense, they do provide a general perspective.

12.2 Some Preliminary Remarks

12.2.1. The cases that are usually considered in mathematical statistics are those which can, in various ways, and in a somewhat loose sense, be regarded as generalizations of the case of exchangeable events. This is in the nature of an aside, however, and simply provides a convenient reference to, and reminder of, the contents of the previous chapter: in fact, in the standard terminology of mathematical statistics one never comes across any reference to exchangeability. In order not to introduce a further difficulty into the task of comparing the various viewpoints, we shall conform to standard usage in this regard. Before doing so, however, we offer the following preliminary clarification of the approach we shall adopt.

We have seen (in Chapter 11, Sections 11.3 and 11.4) that the notions of exchangeability and partial exchangeability can be reduced to that of conditional independence (albeit sometimes in a merely formal sense). More precisely, we have seen that the probability distribution in such cases is always a mixture (i.e. a convex linear combination) of distributions representing independence. If to each case of independence there corresponds an objectively defined 'hypothesis' – like, for example, the proportion of white balls in an urn of unknown composition – then the mixture has an objectively meaningful interpretation. Where this is not the case, the representation is merely formal (as, for example, in the case of a biased coin). There is, however, no difficulty – apart from that of a conceptual nature – in dealing with such 'hypotheses' in these cases as if they were objectively meaningful: for example, one might refer, quite improperly, to 'the hypothesis that the unknown probability of obtaining Heads with the bent coin has the value p' On the other hand, this pseudo-interpretation could always be treated as an asymptotic interpretation of a property that can be defined in a finitistic way by referring to 'frequency in a large number of trials' instead of to 'unknown probability'.[1]

[1] We observe that although the present approach may, at first sight, appear to be very similar (or even equivalent) to that based on 'limit-frequency', there is, in fact, a great deal of difference. We in no way assume the existence of any limit to which the frequency Y_n/n must tend (either with certainty, or in some probabilistic sense – like weak, mean-square or strong convergence and so on). Nor do we utilize any probabilistic form of Cauchy convergence (Chapter 6, 6.8.7), even though this holds (see Chapter 11, 11.4.2) under the assumption of exchangeability (it does not define a 'limit random quantity' and, in any case, requires an infinite number of 'trials'). We base ourselves solely on the frequency Y_n/n of successes in the trials actually considered, whatever they may be, and however many there are, and on the fact that their distribution F_n (the probability distribution of Y_n/n according to the evaluation made at the beginning of the trials) provides an approximation (which improves as n increases) to the limit distribution F, whose existence is therefore established (and this is the only thing we need!).

We shall therefore adopt, in line with our previous remarks, the standard practice of talking in terms of 'hypotheses', irrespective of whether these exist objectively, or merely formally (in which case, they might be interpreted, if at all, in the asymptotic sense given above). Within this framework, the Bayesian approach consists in considering an initial distribution of probability among these hypotheses, this distribution being modified as new information becomes available.

12.2.2. The enormous range of possible applications might lead one to expect a large number of different theoretical models. On the other hand, if one thinks in terms of the basic simple forms of representation, the possibilities are more limited. (Indeed, one might argue that from a Bayesian point of view there is only one form of the problem since everything reduces, in the final analysis, to an application of Bayes's theorem.) In fact, those cases which form the bulk of mathematical statistics can probably be reduced to one or other of the two forms already mentioned: exchangeability or partial exchangeability. There is a meaningful distinction to be drawn between these two cases and, indeed, partial exchangeability embraces a wide range of possible deviations from the exchangeable case. Moreover, within the two categories there are a number of problems of detail and, depending on the field of events under consideration, these may present various levels of difficulty (without there necessarily being any great conceptual difficulties).

Exchangeability and partial exchangeability can also arise in the context of multi-events in general (as we mentioned in Chapter 11), as well as for vectors (r-tuples of random quantities), functions,..., and random elements of any space whatsoever.

What is important in these cases is not so much their actual form, or that of the space to which they belong, but rather the kinds of 'hypotheses' that are assumed; that is the corresponding distributions. There are three main distinctions worth making in this respect: the *discrete* case (involving a finite or countable set of hypotheses); the *parametric* case (involving a set of hypotheses which can each be represented in terms of a fairly restricted set of parameters; i.e. by a vector in a parameter space whose dimension is not too large); *the nonparametric case* (where either the individual hypothesis cannot be represented in terms of a finite number of parameters, or, alternatively, the number of parameters involved is prohibitively large).

Similar distinctions can be made in the case of forms of representation required for *partial* exchangeability (and we shall shortly give a more precise account of these).

12.2.3. In presenting the mathematical development of these ideas, we shall normally deal with problems involving random quantities in the parametric case; in particular, with just one parameter. This is the most straightforward and meaningful case, and hence the most convenient for illustrating the mathematics. It will be immediately obvious, however, that our treatment is completely general, provided the expressions given and the comments made are interpreted in an appropriate manner.

Specifically – to use what we regard as the correct terminology – we shall be dealing with a collection of exchangeable random quantities X_h (see Chapter 11, Sections 11.3 and 11.4). These can be represented in precisely the same way as we saw earlier in the case of exchangeable events: in other words, they can, in a formal sense at least, be thought of as 'independent conditional on some given set of hypotheses'. More precisely, each 'hypothesis' indexes a distribution and we assume that, conditionally, 'all the X_h have this same distribution, and are stochastically independent of one another'.

This implies that their joint distribution is a mixture of products of individual factors (corresponding to the case of independence). As we remarked earlier on, we shall take this representation as our starting point; the interpretation in terms of exchangeability then becomes merely a preliminary clarification.

We shall follow the usual practice in mathematical statistics and work in terms of probability densities (for a justification of this, see the remarks in Section 12.4.3). Expressed mathematically, our basic assumptions then become the following:

- there exists a set of 'hypotheses', a general element of this set being denoted by θ (a point in the hypothesis space); in particular, we shall first consider the case where each hypothesis can be represented by a single real-valued parameter θ;
- conditional on each hypothesis θ (i.e. on the value of θ for the case in question), all the X_h have exactly the same distribution, that is the same density $f(x|\theta)$, and are stochastically independent; this implies that the joint density $p^m(x^1,\ldots, x^m|\theta)$[2] for m of the X_h (no matter how they are chosen or labelled) is given by the product of the densities

$$p^m\left(x^1,x^2,\ldots,x^m\,|\,\theta\right)=f\left(x^1\,|\,\theta\right)f\left(x^2\,|\,\theta\right)\ldots f\left(x^m\,|\,\theta\right);$$

- over the set of hypotheses we have prior probability with density $\pi_0(\theta)$; in the case we are considering, we have a non-negative function of θ such that

$$\int_{-\infty}^{\infty}\pi_0(\theta)\mathrm{d}\theta=1.$$

We note that this latter assumption is the hallmark of the Bayesian approach, whereas other approaches attempt to do without it. We shall develop our treatment within the Bayesian framework but, as we proceed, we shall discuss the techniques that are used by those who eschew the use of 'prior probabilities'.

It follows immediately from these assumptions that the marginal (prior) distributions for any individual X_h, or for m of them, are, expressed as densities, given by

$$f_0(x)=\int f(x\,|\,\theta)\pi_0(\theta)\mathrm{d}\theta, \tag{12.1}$$

$$\begin{aligned} p_0^m\left(x^1,x^2,\ldots,x^m\right)&=\int p^m\left(x^1,x^2,\ldots,x^m\,|\,\theta\right)\pi_0(\theta)\mathrm{d}\theta\\ &=\int f\left(x^1\,|\,\theta\right)f\left(x^2\,|\,\theta\right)\ldots f\left(x^m\,|\,\theta\right)\pi_0(\theta)\mathrm{d}\theta. \end{aligned}\tag{12.2}$$

Here, and elsewhere, it is to be understood that the integrals are to be taken over the entire range of the distribution (and there is no harm in thinking of this as the whole real line, since any range where the density is zero will give a zero contribution). Note that, if we interpret the quantities involved in an appropriate manner, these expressions apply equally well to any abstract spaces (and, in particular, to the case of several parameters, where θ represents a vector).

From the point of view of interpretation, note that f_0 and p_0^m give the previsions of f and p^m if the latter are considered as functions of the random quantity θ. Also note that

[2] The reason for using superscripts will become clear in Section 12.2.5. Note that f is a special case of p^m for $m=1$ ($f=p^1$, and, in what follows, $f_n=p_n^1$, etc.). Usually, however, we shall use f in preference to p^1 in order to make the case $m=1$ more immediately distinguishable, and also to avoid the superscript.

p_0^m, like p^m, is a symmetric function of the x^h (in line with our original assumption of exchangeability).

12.2.4. It is equally straightforward to derive expressions, similar to the above, for the evaluations conditional on knowledge of the values of any of the X_h, or of any n of them. We shall denote these random quantities by X_1, or $X_1, X_2, \ldots, X_n$, partly for convenience and partly to fix ideas for the case in which we observe them in chronological order (and although this might be useful, the reader should remember that it is not an essential part of the argument). The choice of which particular X_h (or set of them) we are interested in calculating the conditional evaluation for is equally irrelevant and the reader should again realize that we denote these by

$$X_{n+1}, X_{n+2}, \ldots, X_{n+m}$$

purely for convenience.

We shall see, in fact, that the evaluations conditional on the knowledge of the values of the first n of the X_h (which we shall denote by f_n and p_n^m) can be expressed in essentially the same form as the f_0 and p_0^m above (the special cases corresponding to $n = 0$; i.e. the initial evaluations, prior to any knowledge of the X_h). In fact, it turns out to be sufficient to determine the distribution $\pi_n(\theta \mid x_1, x_2, \ldots, x_n)$[3] for the parameter θ, conditional on the values $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$, and to substitute this in place of $\pi_0(\theta)$ in the expressions for f_0 and p_0^m. Let us first see this for the case $n = 1$.

After having observed the value of any one of the X_h, $X_1 = x_1$, say, the probability distribution of the parameter (or, more accurately,[4] the distribution 'conditional on the hypothesis $X_1 = x_1$') becomes

$$\begin{aligned} &\pi_1(\theta \mid x_1) = K\pi_0(\theta) f(x_1 \mid \theta) \\ &\left(\frac{1}{K} = \int f(x_1 \mid \theta)\pi_0(\theta)\mathrm{d}\theta = f_0(x_1)\right). \end{aligned} \tag{12.3}$$

This is a straightforward application of Bayes's theorem, or, if one prefers, it is sufficient to observe that the joint density for (θ, X_1) is given both by $\pi_0(\theta) f(x_1 \mid \theta)$ and by $f_0(x_1)\pi_1(\theta \mid x_1)$.

It follows immediately that

$$f_1(x \mid x_1) = \int f(x \mid \theta)\pi_1(\theta \mid x_1)\mathrm{d}\theta, \tag{12.4}$$

$$p_1^m\left(x^2, x^3, \ldots, x^{m+1} \mid x_1\right) = \int p^m\left(x^2, x^3, \ldots x^{m+1} \mid \theta\right)\pi_1(\theta \mid x_1)\mathrm{d}\theta. \tag{12.5}$$

12.2.5. Before we go on with our development, we need to make a comment about notation. The use of superscripts for the x (x^h rather than x_h) is necessary in order to distinguish the use of the values x^h of X_h as 'names of coordinates' for the distribution of X_h (as yet unknown, or considered as unknown), from the use of the x_h as observed values (or values

[3] N.B. For the sake of brevity, we shall sometimes write this as $\pi_n(\theta)$ omitting any explicit mention of $x_1, x_2, \ldots, x_n$ (which must of course be understood).

[4] More accurately' by virtue of what we said in Chapter 11, 11.2.2 (and what we shall say in Section 6 of the Appendix).

assumed to be known). The practical effect of this can be observed in the formulae, where superscripts precede the vertical bar and subscripts follow it (except when rôles are reversed, as happens during the application of Bayes's theorem; see equations 12.11 and 12.11)). In the case of something like $f_1(x|x_1)$, it is clear that it would be superfluous to write $f_1(x^2|x_1)$, because the superscripts are only useful for distinguishing between the x^h when there is more than one of them. For a single (generic) coordinate, it is sufficient to denote it by x.

12.2.6. The expressions for f_1 and p_1^m (and we recall that the former is a special case of the latter; $f_1 = p_1^1$) can be rewritten in a different form, so as to show up certain interesting features more clearly:

$$\begin{aligned} f_1(x|x_1) &= K\int f(x|\theta)\big[f(x_1|\theta)\pi_0(\theta)\mathrm{d}\theta\big] \\ &\left(\frac{1}{K} = \int[\ldots] = f_0(x_1)\right); \end{aligned} \tag{12.6}$$

$$\begin{aligned} p_1^m\left(x^2,x^3,\ldots,x^{m+1}|x_1\right) &= K\int p^m\left(x^2,x^3,\ldots,x^{m+1}|\theta\right)\big[f(x_1|\theta)\pi_0(\theta)\mathrm{d}\theta\big] \\ &= K\int\prod_{i=2}^{m+1} f\left(x^i|\theta\right)\big[f(x_1|\theta)\pi_0(\theta)\mathrm{d}\theta\big]. \end{aligned} \tag{12.7}$$

In this way, we emphasize the fact that f_1 and p_1^m are mixtures of f and p^m, with 'weights' as given in square brackets. Alternatively, we could remove the separation between factors in x_1 and those in x^i ($i = 2, 3,\ldots, m+1$) and write instead

$$\begin{aligned} f(x|x_1) &= K\int\big\{f(x|\theta)f(x_1|\theta)\big\}\pi_0(\theta)\mathrm{d}\theta \\ &= Kp_0^2(x_1,x) = \frac{p^2(x_1,x)}{f_0(x_1)}, \end{aligned} \tag{12.8}$$

or even

$$f_1(x|x_1) = f_0(x)\frac{f_1(x_1|x)}{f_0(x_1)}. \tag{12.9}$$

In this way, we directly emphasize the interpretation in terms of the theorem of compound probabilities and Bayes's theorem.

Proceeding in a similar fashion, we can derive the more general result

$$p_1^m\left(x^2,x^3,\ldots,x^{m+1}|x_1\right) = \frac{p_0^{m+1}\left(x_1,x^2,x^3x,\ldots,x^{m+1}\right)}{f_0(x_1)}. \tag{12.10}$$

In order to derive a form analogous to that of equation 12.9, that is

$$p_1^m\left(x^2,x^3,\ldots,x^{m+1}|x_1\right) = \frac{p_0^m\left(x^2, x^3,\ldots, x^{m+1}\right)f_m\left(x_1|x^2,x^3,\ldots,x^{m+1}\right)}{f_0(x_1)}, \tag{12.11}$$

we must introduce the f_m for $m > 1$; the result is then immediate.

12.2.7. It would have been perfectly straightforward, and more in line with the approaches more commonly adopted in statistics, to have considered right away the distributions conditional on n values,

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$$

instead of on just one. Our main consideration in starting off with the case $n = 1$ is that it enables one to bring out the fact that the effect of n observations is simply the combined effect of considering them one at a time, rather than some magical consequence of there being enough of them to be 'statistically' relevant.

The probability distribution of the parameter θ, given the observed values $x_1, x_2, \ldots, x_n$, is given by

$$\begin{aligned}\pi_n(\theta \mid x_1, x_2, \ldots, x_n) &= K\pi_0(\theta) p^n(x_1, x_2, \ldots, x_n \mid \theta) \\ &= K\pi_0(\theta) f(x_1 \mid \theta) f(x_2 \mid \theta) \ldots f(x_n \mid \theta),\end{aligned} \tag{12.3$'$}$$

where

$$\frac{1}{K} = \int p^n(x_1, x_2, \ldots, x_n \mid \theta)\pi_0(\theta)\mathrm{d}\theta = p_0^n\left(x_1, x_2, \ldots, x_n\right).$$

As we remarked earlier, it is sufficient to replace π_0 by π_n in order to obtain the distributions for one of the X_h, or for m of them:

$$f_n(x \mid x_1, x_2, \ldots, x_n) = \int f(x \mid \theta)\pi_n(\theta \mid x_1, x_2, \ldots, x_n)\mathrm{d}\theta \tag{12.4$'$}$$

$$= K\int f(x \mid \theta)\left[\prod_{h=1}^{n} f(x_h \mid \theta).\pi_0(\theta)\mathrm{d}\theta\right] \tag{12.6$'$}$$

$$= p_0^{n+1}\left(x_1, x_2, \ldots, x_n, x\right) / \, p_0^n\left(x_1, x_2, \ldots, x_n\right) \tag{12.8$'$}$$

$$= f_0(x).p_1^n(x_1, x_2, \ldots, x_n \mid x) / \, p_0^n(x_1, x_2, \ldots, x_n); \tag{12.9$'$}$$

$$\begin{aligned}&p_n^m\left(x^{n+1}, x^{n+2}, \ldots, x^{n+m} \mid x_1, x_2, \ldots, x_n\right) \\ &\quad = \int p^m\left(x^{n+1}, \ldots, x^{n+m} \mid \theta\right)\pi_n(\theta \mid x_1, \ldots, x_n)\mathrm{d}\theta\end{aligned} \tag{12.5$'$}$$

$$= K\int \prod_{i=n+1}^{n+m} f\left(x^i \mid \theta\right).\prod_{h=1}^{n} f(x_h \mid \theta)\pi_0(\theta)\mathrm{d}\theta \tag{12.7$'$}$$

$$= p_0^{n+m}\left(x_1, \ldots, x_n, x^{n+1}, \ldots, x^{n+m}\right)/p_0^n\left(x_1, \ldots, x_n\right) \tag{12.10$'$}$$

$$= \frac{p_0^m\left(x^{n+1}, \ldots, x^{n+m}\right)p_m^n\left(x_1, \ldots, x_n \mid x^{n+1}, \ldots, x^{n+m}\right)}{p_0^n\left(x_1, \ldots, x_n\right)}. \tag{12.11$'$}$$

The last four expressions include all the others as special cases: more precisely,

equations 12.5′, 12.7′, 12.10′ and 12.11′ give for general n and m, what equations 12.5, 12.7, 12.10 and 12.11 give for $n = 1$ (with m arbitrary), and

equations 12.4′, 12.6′, 12.8′ and 12.9′ give for $m = 1$ (with n arbitrary) and
equations 12.4, 12.6, 12.8 and 12.9 give for $n = m = 1$.

The interpretations are identical to those that we gave in the simplest case (i.e. that of $n = m = 1$) and, when contemplating extensions to cases that are more complicated (insofar as the formulae are concerned, anyway), it may be useful to bear this case in mind.

Given $x_1, x_2, \ldots, x_n$, the likelihoods for θ and x are

a) $\prod_h f(x_h \mid \theta)$ (as a function of θ),

b) $\int f(x \mid \theta) \prod_h f(x_h \mid \theta) \pi_0(\theta) \mathrm{d}\theta$ (as a function of x),

respectively. In fact, any function differing from (a) by a factor independent of θ, or from (b) by a factor independent of x, could be taken as the respective likelihood.

12.2.8. We now turn to the case of 'partial exchangeability'. The account we shall give will be even shorter than the above and we shall rely on the examples to clarify our interpretation and approach to the problem.

In terms of our formulation, this case differs from the previous one in that, conditional on each of the 'hypotheses' characterized by θ, the X_h are still stochastically independent, but now may have different distributions. The latter depend not only on the parameter θ, but also on certain observable quantities y_h, which relate to the X_h. Like θ, y may be real-valued, or a vector, or whatever (irrespective of the form of θ).[5] In order to keep the presentation on a simple level, we shall take y to be real-valued (the general case presents nothing new from a conceptual viewpoint).

Formally, instead of starting from $f(x|\theta)$ we consider $f(x|\theta, y)$. So far as the prior distribution for θ is concerned, nothing changes; we begin, as before, with some $\pi_0(\theta)$. The distribution $p_0^m(x^1, x^2, \ldots, x^m)$ (knowledge of which enables one to derive everything else) will, however, also depend on the values that y takes for each of the $X_x, X_2, \ldots, X_m$, and has the form

$$p_0^m\left(x^1, x^2, \ldots, x^m\right) = \int f\left(x^1 \mid \theta, y_1\right) f\left(x^2 \mid \theta, y_2\right) \ldots f\left(x^m \mid \theta, y_m\right) \pi_0(\theta) \mathrm{d}\theta, \quad (12.12)$$

where y_h corresponds to X_h. If we wish this to be made explicit, we must write the left-hand side as

$$p_0^m\left(x^1, \ldots, x^m \mid y_1, \ldots, y_m\right).$$

On the other hand, if we do make systematic use of this explicit form the expressions become rather cumbersome – particularly those which are already complicated, even without this additional detail.

The following are intended as examples of the kinds of y_h that might be observed[6] and considered as possibly influencing the distribution of X_h: the temperature at the time at

[5] In other words, one could be a vector and the other a real number, etc.

[6] See the remark at the end of Section 12.3.3.

which the experiment yielding X_h took place; the age of an individual whose reaction to some given drug is measured by X_h; the precision of the instrument which performs the measurement giving X_h; and so on. We shall shortly give an example involving the latter possibility.

12.3 Examples Involving the Normal Distribution

12.3.1. Given that the normal distribution is widely used (and somewhat abused) in statistics, it is natural that the most familiar problems of inference are those which involve this distribution. The prevision m and the variance σ^2 suffice to characterize the distribution, which is usually denoted by $N(m, \sigma^2)$. The density, as we already know, is given by

$$f(x)=\frac{1}{\sqrt{(2\pi)}\sigma}\exp\left\{-\frac{1}{2}(x-m)^2/\sigma^2\right\}.$$

By far the most important case is that in which m corresponds to the unknown parameter θ (while σ^2 is known), but we shall also deal with the opposite case ($\theta = 1/\sigma^2$, m known) and with the case in which both parameters are unknown (θ is the 'vector' (θ_1, θ_2), $\theta_1 = m$ and $\theta_2 = 1/\sigma^2$).[7] On the basis of these examples, all under the assumption of complete exchangeability, we can discuss variants corresponding to partial exchangeability. The simplest such variants are obtained by replacing the assumption 'σ^2 known' (or 'm known') by '$=y$', known for each X_h, but possibly different for different h'.

12.3.2. *The case where m is unknown.* This is, above all, the case considered in the theory of errors (experimental or observational) as applied in astronomy geodesy, physics and so on. What is unknown is the *true* value of the quantity that is being measured; that is $\theta = m$. The accuracy of the instrument (as represented by σ^2) is assumed known and the distribution of the observed value is assumed to be $N(m, \sigma^2)$; that is a normal distribution centred at the *true* value and having the given precision.

We have, therefore,

$$f(x\,|\,\theta)=K\exp\left\{-\frac{1}{2}(x-\theta)^2/\sigma^2\right\}, \tag{12.13}$$

and (apart from the constant factor K) this is the likelihood for θ given by an observation x. The likelihood given by n observations $x_1, x_2,\ldots, x_n$ is

$$\prod_h f(x_h\,|\,\theta)=\exp\left\{-\frac{1}{2\sigma^2}\sum_h (x_h-\theta)^2\right\}.$$

[7] In the terminology used in the theory of errors, the reciprocal $1/\sigma$ of the standard deviation is called the *precision*, and the reciprocal $1/\sigma$ of the variance is called the *weight* (although sometimes a different unit of measure is used; e.g. precision $1/\sqrt{2}\sigma$, weight σ_0^2/σ^2, where σ_0^2 is chosen as appropriate for the problem under consideration). It might seem rather unnecessary to have four terms available, but this is not entirely the case.

We shall take the *weight* σ^{-2} as the parameter θ, instead of the more customary variance, σ^2, since this turns out to simplify the formulae.

Noting that

$$\begin{aligned}\sum_h (x_h-\theta)^2 &= \sum_h \left(x_h^2-2x_h\theta+\theta^2\right) = \text{const.} - 2\theta\sum_h x_h + n\theta^2\\ &= n\left(\text{const.} - 2\theta\frac{1}{n}\sum_h x_h + \theta^2\right)\\ &= n\left[\text{const.} + (\bar{x}-\theta)^2\right],\end{aligned}$$

where $\bar{x} = 1/n\sum_h x_h$ the mean of the x_h, we see that the likelihood can be rewritten in the form

$$\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\}. \tag{12.14}$$

In other words, it has the same form as the likelihood of a single observation equal to the mean $\bar{x}$, and with *standard deviation* $\sigma/\surd n$ (i.e. reduced in the ratio 1 to $1/\surd n$, which is equivalent to *precision* increased in the ratio 1 to $\surd n$, *variance* reduced in the ratio 1 to $1/n$, *weight* increased in the ratio 1 to n).

The posterior distribution for θ is therefore given by

$$\pi_n(\theta) = K\pi_0(\theta)\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\right\}. \tag{12.15}$$

Since the likelihood is maximized for $\theta = \bar{x}$ and decreases as we move away from this value (the decrease being sharper for larger n), the posterior distribution concentrates around x. In particular, if the prior distribution is taken to be normal, $N(m_0, \sigma_0^2)$, say then the posterior distribution is also normal. More precisely, the posterior (or final[8]) distribution is $N(m_\text{f}, \sigma_\text{f}^2)$, where

$$m_\text{f} = \frac{m_0\sigma_0^{-2} + n\bar{x}\sigma^{-2}}{\sigma_0^{-2} + \sigma^{-2}}, \qquad \sigma_\text{f}^{-2} = \sigma_0^{-2} + n\sigma^{-2}. \tag{12.16}$$

In words: the posterior *weight* (1/*variance*) is the sum of the weights from the prior and the likelihood; the posterior mean is the weighted mean of the prior mean and the mean from the likelihood (m_0, and n times $\bar{x}$; i.e. a function of m_0 and $x_1, x_2,\ldots, x_n$), the weights being the respective *weights* (thus revealing the aptness of the terminology).

12.3.3. The extension to the case of 'observations made with different precisions' is immediate. Let us assume, for instance, that we know that the n observations are performed with different measuring instruments, the errors of which have standard deviations $\sigma_1, \sigma_2,\ldots, \sigma_n$. It is clear that (by an argument similar to that used above) these observations are equivalent to a single observation whose value is given by the weighted mean of the x_h (with weights σ_h^{-2}) and having weight equal to the sum of the weights.

[8] *Translators' note.* The terms *prior* and *posterior* seem firmly established in English publications relating to applications of Bayes's theorem, and we have used them in preference to the terms *initial* and *final*. The Italian version uses the latter, and the notation m_f and σ_f^2 derives from this usage.

If the prior distribution is $N(m_0, \sigma_0^2)$, the posterior distribution is given by $N(m_f, \sigma_f^2)$, where m_f and σ_f^2 are determined by the weighting process just described, except that we now also include m_0 with weight σ_0^{-2}.

This is an example of 'partial exchangeability' with $y_h = \sigma_h^2$ (or we could take $y_h = \sigma_h^{-2}$, and

$$f(x\,|\,\theta, y) = K\exp\left\{-\frac{1}{2}(x-\theta)^2 / y\right\}.$$

We should draw attention to the fact that the y_h must actually be known and observed for each X_h under consideration. In our example, we must know with what precision each measurement has been performed. One should be careful not to think of it as being sufficient to know that each measurement has been performed using instruments of various precisions (for example, by choosing each time at random from among some given collection of measuring instruments, but without registering which were actually used and how often). Under this latter assumption, one would have a case of exchangeability with

$$f(x\,|\,\theta) = \sum_k c_k f_k(x\,|\,\theta), \quad f_k(x\,|\,\theta) = K\exp\left\{-\frac{1}{2}(x-\theta)^2 y_k\right\}, \quad c_k \geqslant 0, \quad \sum_k c_k = 1$$

(i.e. no longer a normal distribution but a mixture of normals). In the same way, in the other examples it would be necessary to have actually noted the temperature, age etc., in each case.

12.3.4. *Comments*. The choice of the normal form for the prior distribution in the case just considered is convenient in that the posterior distribution is then always a member of this same family. We shall see that in other cases, too, we can find distributions for which this property holds.

On the other hand, this convenience does not justify our making such a choice if it is not compatible with our actual prior opinion; neither does it provide any *a priori* justification for regarding such distributions as in any way playing a special role. A reasonable approach involves adopting 'convenient' distributions if and insofar as they provide a sufficiently accurate representation of one's actual opinions (and this is especially useful in those problems where the precise form chosen has little influence on the final outcome).

If the influence on the final outcome is going to be practically negligible, one might even 'omit' the factor $\pi_0(\theta)$ altogether; that is, to be more precise, one might consider the limit case of a 'constant density'. This improper distribution could be interpreted, for example, as the limit of the normal distribution $N(0, \sigma^2)$ as $\sigma^2 \to \infty$, or of the uniform distribution

$$\pi_0(\theta) = \frac{1}{2}a\left(-a \leqslant \theta \leqslant a\right) \text{ as } a \to \infty.$$

As we shall see, other forms of improper prior distribution may be more appropriate, depending on the form of the problem.

Sometimes the use of the improper, uniform prior distribution is interpreted as representing 'total ignorance'. This is nonsense: every distribution reflects some sort of opinion, and none of these have any special status – not even in the negative sense of representing no opinion at all. Moreover, one should note that the uniform distribution is not invariant under changes in the parametrization (e.g. θ into $\log\theta$ or e^{θ} etc.). A number of useful observations of this kind can be found in Lindley, Vol. II (with particular reference to this topic, see p. 145).

Remarks. The above considerations are all dependent on a certain mathematical point which should be clearly understood, because it serves to clarify the particular practical consequences of the above.

Rigorously speaking, a *density is not just a point function but rather a function of the point and of the measure* assumed over the space under consideration. For instance, it is well known that in the case of measures defined in terms of coordinate systems a change of coordinates alters the density by multiplying it by the Jacobian (and the same thing holds more generally). It follows, for instance, that we could always arrange to have a constant density (it suffices to take the distribution corresponding to such a density as the underlying measure).

The *likelihood*, on the other hand, actually is a point function, and 'equating' it to a density is a meaningless idea. We can always achieve what we want, however, by an appropriate choice of the measure (which is never significant from a theoretical viewpoint), taking it over the most convenient reference system (or one which is sufficiently convenient) in order to make calculations as straightforward as possible.

We shall, therefore, find it useful (and a number of examples of this will be given) to choose a family of prior distributions with density 'equal' to the likelihood. In the terminology introduced by Raiffa and Schlaifer, these constitute the *conjugate* family for the problem. One should note, however, that this notion has no absolute meaning, but can be useful relative to some given standard formulation of a problem.

12.3.5. *The case where σ^2 is unknown.* We again consider the normal distribution, $N(m, \sigma^2)$ but now with the variance σ^2 unknown (and it is convenient to set $\theta = 1/\sigma^2 =$ 'weight') and the mean m known. This case arises, for example, if one wishes to calibrate a new measuring instrument (i.e. to determine its precision as measured by $\theta = 1/\sigma^2$) by making repeated measurements of a given known quantity m.

In this case we have

$$f(x\,|\,\theta) = K\theta^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\theta(x-m)^2\right\}. \tag{12.17}$$

As a function of θ (and leaving aside factors independent of θ), this is the likelihood for θ given by an observation x.

The likelihood for θ given by n observations $x_1, x_2, \ldots, x_n$ is, therefore,

$$\prod_h f(x_h\,|\,\theta) = \theta^{\frac{1}{2}n} \exp\left\{-\frac{1}{2}\theta\sum_h (x_h-m)^2\right\} = \theta^{\frac{1}{2}n} e^{-\frac{1}{2}\theta S^2}, \tag{12.18}$$

where $S^2 = \sum_h (x_h - m)^2$ (the constant K^n having been omitted).

Since the form of this expression is that of the density of a gamma distribution, any choice of prior from within the gamma family will ensure that the posterior distribution belongs to the same family (and the comments of Section 12.3.4 should be understood in this case, too). Taking

$$\pi_0(\theta) = K\theta^{\alpha-1}\mathrm{e}^{-\lambda\theta},$$

we obtain

$$\pi_0(\theta) = K\theta^{\alpha-1+\frac{1}{2}n}\mathrm{e}^{-\left(\lambda+\frac{1}{2}S^2\right)\theta}. \tag{12.19}$$

12.3.6. *The case where both m and σ^2 are unknown.* This arises in the context of errors of observation, as in Section 12.3.2, except that we also assume the precision of the measuring instrument to be unknown. It is also the most frequently studied case in statistics, where we have a population (of individuals, objects, experiments etc.) in which some given quantity (X_h for the *hth* individual) is known (or assumed) to be normally distributed, but with neither the mean (central) value nor the variance known.

The example involves two parameters – those encountered separately in the previous two cases. We put $\theta_1 = m$, $\theta_2 = 1/\sigma^2$, and hence we obtain

$$f(x \mid \theta_1, \theta_2) = K\theta_2^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\theta_2 (x - \theta_1)^2\right\}. \tag{12.20}$$

The likelihood for θ_1 and θ_2 after having observed $x_1, x_2, \ldots, x_n$ is given by

$$\theta_2^{n/2} \exp\left\{-\frac{1}{2}\theta_2 \sum_h (x_h - \theta_1)^2\right\} = \theta_2^{n/2} \exp\left\{-\frac{1}{2}\theta_2 \left[\nu s^2 + n(\bar{x} - \theta_1)^2\right]\right\}, \tag{12.21}$$

where

$$\nu = n - 1, \quad s^2 = \sum_h (x_h - \bar{x})^2 / \nu, \quad \bar{x} = \sum_h x_h / n$$

(the steps are the same as those given in Section 12.3.2, except that the constant νs^2 can now no longer be omitted because it is multiplied by the parameter θ_2).

The standard assumption for the prior distribution is, in this case, the improper uniform distribution for both θ_1 and $\log \theta_2$: this results in an improper 'density proportional to $1/\theta_2$'

$$\left(\text{over the half-plane} - \infty < \theta_1 < +\infty, 0 < \theta_2 < +\infty\right).$$

Strictly speaking, this assumption is only made by Bayesians (and even then only by those who have no objections to improper distributions), but there is some justification for referring to it as the standard assumption because it leads to the same conclusions as those arrived at by non-Bayesian statisticians using other methods of approach.

With these assumptions, that is supposing that we are prepared to express the prior density in the form

$$\pi_0(\theta_1, \theta_2) = K/\theta_2, \tag{12.22}$$

we obtain

$$\pi_n\left(\theta_1,\theta_2\right)=K\theta_2^{(n-2)/2}\exp\left\{-\frac{1}{2}\theta_2\left[\nu s^2+n\left(\bar{x}-\theta_1\right)^2\right]\right\}. \tag{12.23}$$

The marginal posterior densities for θ_1 and θ_2 (obtained by integrating out the other variable) are[9]

$$\begin{gathered}\pi_n^{(1)}\left(\theta_1\right)=K\left\{1+n\left(\bar{x}-\theta_1\right)^2/\nu s^2\right\}^{\frac{1}{2}n}=K\left(1+t^2/\nu\right)^{-\frac{1}{2}(\nu+1)}\\ \left(t=\sqrt{}n\frac{\bar{x}-\theta_1}{s}\right),\end{gathered} \tag{12.24}$$

$$\pi_n^{(2)}\left(\theta_2\right)=K\theta_2^{\nu/2}\exp\left\{-\frac{1}{2}\nu s^2\theta_2\right\}. \tag{12.25}$$

For θ_2, we still have a gamma distribution, just as in the case m known (Section 12.3.5). For θ_1, on the other hand, the normal distribution, which we obtained in the case σ^2 known (Section 12.3.2), has been replaced by Student's distribution (see the very end of Chapter 10). The effect of not knowing m and σ^2 is that they are replaced by $\bar{x}$ and s^2 (which are 'reasonable estimates' of them), and that, in the case where we are ignorant of σ^2, the normal is replaced by the Student distribution which has much fatter tails (although it tends to the normal as $n \to \infty$). For large n, therefore, the difference is practically negligible.

12.4 The Likelihood Principle and Sufficient Statistics

12.4.1. Given that we started out by adopting the Bayesian approach and that we have adhered to it coherently throughout, the 'likelihood principle' inevitably appears to be rather obvious and certainly not worth getting excited about. It simply states that the information available from any set of observations is entirely contained in the corresponding likelihood function. Since this is, in fact, the factor which transforms the prior opinion into the posterior, this is all we require and, indeed, all we can ask for.

If, however, one proposes ad hoc methods – more or less on a trial and error basis – it might well happen that they conflict with this 'principle'. For this reason, non-Bayesians have debated among themselves as to whether this principle (or a variant thereof) should be rejected, or whether, on the contrary, one should reject methods which do not comply with it (or whether such methods could be considered valid as approximations).

From the Bayesian standpoint there are two possible reasons for wishing to mention the principle: firstly, to warn against superficial interpretations of it; secondly, as a starting point for developing the topic of 'sufficient statistics'.

[9] For the details of the calculations, see, for example, Lindley, 5.3 and 5.4, but note that he takes $\theta_2=\sigma^2$ (whereas by setting $\theta_2=1/\sigma^2$, we have obtained a certain amount of simplification in the formulae and calculations).

The warnings are the obvious ones (but, on the other hand, mistakes are often the result of overlooking the obvious) and concern a too literal interpretation of the following statement of the principle:

> *The information contained in a set of observations is completely summarized by the likelihood function, and provided it can be combined with similar results etc., it is quite sufficient to quote this.*

All is well, provided we also enter the following reservation: '*so long as the basic assumptions remain unchanged*'. If, for instance, one starts off by assuming that certain errors 'are normally distributed' and then begins to wonder whether they, in fact, have some other distribution, the data expressed in the form of the 'likelihood function' can no longer provide all relevant information.

The discussion about '*sufficient statistics*' could begin by our pointing out that the likelihood function is itself a sufficient statistic (that is to say, it provides an exhaustive summary of the information contained in the data). It follows, therefore, that the summaries of the data that characterize the likelihood, when considered altogether, themselves form a sufficient statistic. In the cases that we have examined, for example, we have the following:

Section	**Case**	**Sufficient statistic**	
12.3.2	(m unknown, σ^2 known and constant)	the pair	$n,\ \bar{x}$;
12.3.3	(m unknown, σ^2 known but varying)	the pair	$\sum y_h, \sum y_h x_h$;
12.3.5	(σ^2 unknown, m known and constant)	the pair	$n,\ S^2$;
12.3.6	(m and σ^2 both unknown)	the triple	$n,\ \bar{x},\ s^2$.

12.4.2. For the sake of completeness, we now give a few of the basic notions relating to the concept of a sufficient statistic.

In what follows, it will be convenient to denote the *data* (which in general consists of n observations, x_h; $h = 1, 2,\ldots, n$) simply by x; the *parameters* (no matter whether there is just one, θ, or several, θ_i, $i = 1, 2,\ldots, s$) by θ; and the *sufficient statistic* (consisting either of a single real-valued function of the data, $t = t(x)$, or of several; $t_j = t_j(x)$; $j = 1, 2,\ldots, r$) by t; in the same way $p(x|\theta)$ denotes something of the form $p^n(x_1, x_2,\ldots, x_n|\theta)$ and so on. From a conceptual point of view, the argument is precisely the same, whether $x, \theta, t,\ldots$ are real numbers, or vectors, or whatever.

Expressing our comments about sufficient statistics in the form of a definition, we have the following (also known as the 'sufficiency principle'): $t(x)$ is a sufficient statistic for the family $p(x|\theta)$ if and only if, for any prior $\pi_0(\theta)$, the posterior distribution is the same no matter whether we condition on x or on $t(x)$; that is. $\pi(\theta|x) = \pi(\theta|t(x))$.

A necessary and sufficient condition for $t(x)$ to be a sufficient statistic for $p(x|\theta)$ is that the latter can be written as

$$p(x|\theta) = f(t(x),\theta).g(x), \qquad (12.26)$$

where *f* and *g* are arbitrary functions. The necessity of the condition is obvious. From the definition, it follows that

$$p(x\,|\,\theta)=p\big(t(x)\,|\,\theta\big)p\big(x\,|\,t(x),\theta\big),$$

and this is then equal to

$$p\big(t(x)\,|\,\theta\big)p\big(x\,|\,t(x)\big).$$

If we now take the first factor as *f* and the second as *g* this is in the required form. The sufficiency part is known as *Neyman's factorization theorem.*

What we have stated above is true in the general case (i.e. *x* does not necessarily have to consist of the results of 'independent observations from the same distribution'). If we go back to the special case (complete exchangeability), we can pose some further problems. In this case, we may be interested in knowing, for example, whether, as *n* varies, we can always obtain a sufficient statistic of fixed dimension (for example, having *r* components: $t=(t_1(x), t_2(x),\ldots,t_\mathrm{r}(x))$).

Ignoring the finer points, the condition for this to happen is that the family of distributions $f(x|\theta)$ is a member of the *exponential family*: that is that it is of the form

$$f(x\,|\,\theta)=F(x)G(\theta).\exp\left\{\sum_{j=1}^{r}u_j(x)\Phi_j(\theta)\right\}, \tag{12.27}$$

where F, G, u_j, Φ_j, are arbitrary functions. In this case, a sufficient statistic, given any *n* observations $x=(x_1, x_2,\ldots, x_n)$, is provided by the *r* functions

$$t_j(x)=\sum_{i=1}^{n}u_j(x_i) \qquad (j=1,2,\ldots,r), \tag{12.28}$$

together with *n* (although the latter is sometimes left to be understood).
In this case, the likelihood function (for θ, on the basis of the given *x*) is

$$p(x\,|\,\theta)=K\ .\ G(\theta)^n\exp\left\{\sum_{j=1}^{r}\ t_j(x)\Phi_j(\theta)\right\}. \tag{12.29}$$

If the prior distribution $\pi_0(\theta)$ is proportional to the form

$$G(\theta)^a.\exp\left\{\sum_{j=1}^{r}b_jf\ _j(\theta)\right\},$$

then the posterior distribution will also have this same form:

$$\pi(\theta\,|\,x)=K\pi_0(\theta)p(x\,|\,\theta)=KssG(\theta)^{a+n}.\exp\left\{\sum_{j=1}^{r}\big[b_j+t_j(x)\big]\Phi_j(\theta)\right\}. \tag{12.30}$$

This explains how it is always possible to define *conjugate families* of distributions whenever we are dealing with a member of the exponential family (and the same advantages are obtained as we saw previously for the normal and gamma distributions).

Recall, however, that the concept lacks any genuine substantial foundation, as we explained in the final paragraph of Section 12.3.4.

12.4.3. The time has now come for us to explain – in line with what we said in Section 12.2.3 – why we have restricted ourselves to cases in which a probability density exists.

Firstly, of course, it is quite natural to restrict oneself to the most straightforward and meaningful practical cases; these are the ones we have mentioned: either the discrete cases or those where a density exists.[10]

Over and above this, however, it is necessary to point out a far more essential reason; one which I do not think I have heard put forward before, nor have had occasion to mention myself. In order for inferences to be valid independently of the indeterminism that arises on account of 'probabilities conditional on events of zero probability', together with related questions concerning 'nonconglomerability' (Chapter 4, Sections 4.18 and 4.19, and Chapter 6, 6.9.5), it is necessary to confine oneself to problems that can be dealt with by using only probabilities conditional on hypotheses having nonzero probability. This happened trivially in the case of 'concentrated masses' and it happens directly in cases where a density exists, provided one assumes – as, fortunately, seems to be 'inevitable' from an empirical point of view – that knowledge of observed values x_h is not 'exact', but that, at best, it involves 'belonging to a neighbourhood of $(x_1, x_2, \ldots, x_n)$' small enough to make it possible to argue in terms of a density, but not in terms of the point itself.

12.5 A Bayesian Approach to 'Estimation' and 'Hypothesis Testing'

12.5.1. The natural way to present the solution of any problem of statistical inference is to give the relevant probability or probability distribution. In the cases we have considered, this involved the posterior distribution given the observed data. Unfortunately, however, such a solution cannot be regarded as 'natural', insofar as it is not 'familiar' to most people. It is for this reason, perhaps, that attempts have been made to replace the posterior distribution with some sort of crude summary conveying a more immediate message.

Two such crude approaches to summarizing the distribution of a random quantity X are widely used: the first consists in providing a unique value $\hat{x}$, around which the distribution is concentrated; the second in providing an interval $[x', x'']$, enclosing a large proportion of the distribution. These descriptions are rather vague but they can be made more precise in various ways and, in so doing, we obtain the various methods of *estimation*. More specifically, in the first case we refer to $\hat{x}$ as a *point estimate*, while in

[10] Or even mixed cases; a distribution admitting a density, plus a few 'concentrated masses' at particular values of interest: for example, the percentage of some given compound in an alloy when the value zero (the absence of the compound) can occur with non-zero probability.

A typical example is the problem of King Hiero's crown (to which the episode of Archimedes' 'Eureka!' refers). Was there silver in the crown? (See L.J. Savage *et al., The Foundations of Statistical Inference*, London, Methuen (1961).) Another example is given by the correlation between two genes (0 corresponds to their being on separate chromosomes).

the second we call $[x', x'']$ an *interval estimate*. Similar considerations apply in higher dimensions (where the form of the 'interval' may be much more general).

There are other cases in which one poses the inferential question in a different way, but where one requires solutions formally similar to those given above. It may be that there is a value x_*, or an interval $[x_*', x_*'']$, for which we wish to know whether or not X is equal to x_* (either exactly or approximately), or whether or not it lies between x_*' and x''. In such cases, one refers to *tests of hypotheses*, because an answer of either YES or NO is required in relation to the so-called '*null hypothesis*'

$$\left(X = x_*, \text{or } x_*' \leqslant X \leqslant x_*''\right).$$

The contrary hypothesis, or the various hypotheses into which the complement may be divided, are known as '*alternative hypotheses*'.

12.5.2. The traditional approach to these problems, and still the most popular, is based on ad hoc methods, which, in contrast to Bayesian methods (based on a systematic and coherent theory), are largely rule-of-thumb.

In the present context, we wish to examine the extent to which they can be modified to fit into the Bayesian framework. In other words, we shall consider them not as separate and distinct methods leading to an alternative set of techniques but rather as useful summaries of certain aspects of the actual, complete solution – that is the description of the posterior distribution.

In certain respects, it is clear that the solution will depend on some value relating to the (posterior) distribution; for example, the prevision (for a fair bet), or some other mean, or the median and so on. Such a value might well be referred to as the (point) estimate for the problem; that is the appropriate *mean* in the Chisini sense.

In many cases, it is clear that giving an interval in which the random quantity of interest might plausibly be thought to lie is more informative than any attempt at actually pin-pointing it. From the Bayesian standpoint, we would give an interval having some stated probability of containing X (usually a high probability; e.g. 95%, 99%: in general, 100β%). In such a case, following Lindley, we could refer to this interval $[x', x'']$ as a 100β% (*Bayesian*) *confidence interval for X*. The qualification 'Bayesian' will be implicit in what follows and the reader should note that a '100β% confidence interval' is a very different concept in a non-Bayesian context (as we shall see), and that it is important to distinguish between the two.

In general, there are infinitely many such intervals for any given level. The standard procedure is to choose the shortest one (in a certain sense, it is the most informative). In many cases – for instance, those for which the density has a unique maximum and decreases on either side of it – this interval is characterized by the fact that at each point inside it the density is greater than at every point outside it. One should note, however (in order that this criterion should not appear more 'natural' than it actually is), that both the length and the density change, in general, if X is transformed into some function of itself. For example, if $[x', x'']$ is a 95% confidence interval for X, the interval $[e^{x'}, e^{x''}]$ remains such for e^X, but if the former is the interval of shortest length, the latter, in general, is not.[11]

[11] This is an obvious consequence of what we have seen more generally (see *Remark*, Section 12.3.4).

12.6 Other Approaches to 'Estimation' and 'Hypothesis Testing'

12.6.1. Those who reject the Bayesian approach cannot base their inferences on the posterior distribution even if they wished to – it does not make any sense so far as they are concerned. As a result, they are forced to have recourse to ad hoc criteria and, hence, to open the floodgates to arbitrariness. This has led to an enormous proliferation of such techniques. For the sake of completeness, and to provide a basis for certain critical comparisons, we shall give a short account of the most important and best known of these.

The basic reason why non-Bayesians are unable to refer to the posterior distribution lies in their rejection of the use of a prior distribution.[12] The best they can then do is to base themselves on the likelihood function; failing that, they simply resort to playing with formulae that are without any real foundation.

The situation can be summarized as follows.

A method for obtaining a point estimate $\hat{x}$ given the data x_h ($h = 1, 2,\ldots, n$), reduces in the final analysis, to providing a formula which expresses $\hat{x}$ as a function of the x_h: $\hat{x} = \phi_n(x_1, x_2,\ldots, x_n)$. (The same thing applies to finding the end-points x' and x'' for an interval estimate.) At the very beginning the choice of the criterion consists in defining a random quantity

$$\hat{X} = \phi_n\left(X_1, X_2, \ldots, X_n\right),$$

a function of the $X_{h,}$ whose value $\hat{X} = \hat{x}$ is to be taken as the estimate.

The problem must always be interpreted as follows (and we express it in a form which should be sufficiently vague to be acceptable to everyone): the X_h are either approximate measurements of some 'true value' x_0, which we would like to know, or they are the values of some given quantity as observed in a sample, and the value x_0 which we wish to know is some typical value (the mean, median, mode …) of the distribution of that quantity in the population. We seek to 'estimate' x_0 by $\hat{x}$.

12.6.2. There are, essentially, three different levels at which this problem can be formulated and dealt with.

At the very lowest level one simply ignores the probabilistic nature of the problem (or, at least, it is not taken into account in the formulation). At this level, we can only examine the formal properties of the proposed function and judge on empirical grounds the extent to which these are appropriate. It is rare to find this approach adhered to in any systematic way but considerations of this kind do crop up incidentally now and again

[12] The paper by B. de Finetti and L.J. Savage, 'Sul modo di scegliere le probabilità iniziali', which we have already quoted several times (see Chapter 11, footnote to Section 11.1.1), and my talk at the Saltzburg conference in 1968, published as B. de Finetti, 'Initial probabilities: A prerequisite for any valid Induction', *Synthese*, XX, 1 (1969), are devoted to a refutation of this, and to the clarification of various problems connected with it.

Related topics were mentioned at the conference by Vetter, Hintikka, von Kutschera and Frey; in particular, see the paper by I.J. Good, 'Discussion of Bruno de Finetti's paper', which reveals the differences in attitudes existing within the subjectivistic conception.

(and there have been attempts to put forward abstract theories of 'methods of measurement' at this level).

The methods proposed by objectivistic statisticians are at an intermediate level. The probabilistic framework is accepted for that which takes place conditional on certain given hypotheses, but any reference to a probability distribution for the hypotheses themselves is rejected. To relate this to our previous considerations, the 'hypotheses' are the various values of the parameter θ and what is rejected is the prior distribution $\pi_0(\theta)$ (and hence the posterior $\pi_0(\theta|x)$).[13] All that one is permitted to work with is the assumption that the X_h are stochastically independent with the same distribution, $f(x|\theta)$, conditional on each value of θ.[14]

The implication of this for problems of estimation (and similarly for 'tests of hypotheses') is that the function ϕ can only be made to depend on the $f(x|\theta)$, whereas in the unrestricted (i.e. Bayesian) formulation one must also make it depend on $\pi(\theta)$.

12.6.3. One way of avoiding the difficulty is to use the Bayesian approach (either consciously or unconsciously) but omitting $\pi(\theta)$: in other words, by implicitly adopting the (possibly improper) prior $\pi(\theta) = \text{constant}$. In this way, the conclusions obtained are necessarily valid, although it should be noted that indiscriminate use of this prior may result in its adoption in situations where neither the individual using it, nor the majority of other people, find it reasonable. Worst of all, actual contradictions can arise if the approach is used independently in related problems (as a trivial example, taking first θ to have a uniform prior and then, later in the same problem, taking $1/\theta$ to be uniform).

One way of running head on into the difficulty – whilst claiming at the same time to have solved the problem – is to assert that nothing can be said concerning the probability of the statement of interest being true (e.g. that x_0 is 'close to $\hat{x}$', or that it lies between x' and x''). Having decided against overcoming this problem by the use of prior probabilities, it suffices … to pretend that the solution we require is, in fact, a different one, and concerns the probability of the statement of interest being true conditional on the (false!) hypothesis that x_0 is known. In fact, the statements would be similar in appearance only. We could gloss over it by saying, in either version, that '*in any case*, it is almost certain that x_0 and $\hat{x}$, the true and estimated values respectively, are close to one another', as if the phrase 'in any case' had some abstract and absolute meaning, both when it refers to 'whatever the true value might be' and when it refers to 'whatever the estimated value might be'.

The fallacy in confusing the two cases is obvious. In fact, under the usual assumptions, it is almost certain that the mean of the measurements obtained from n observations will turn out to be near the true value (whatever it may be), since we are assuming the error distribution to be the same no matter what the true value is. When we consider the mean resulting from a set of given observed values, however, we are by no means entitled to conclude that the true value is almost certainly near this mean. It may well be that, finding the latter conclusion hard to believe, one considers it as much more

[13] There is no objection, however, in problems where objectivists adjudge there to be an 'objective' prior. In such cases, the approach will be the same as a Bayesian would adopt.

[14] There seems little point in complicating this brief account by extending it to include the more general cases (e.g. those with $f(x|\theta, y)$, and so on).

plausible (even though *a priori* quite improbable) that, by chance, the observations have turned out to be affected by large errors acting in that particular direction.[15]

12.6.4. There are critics who occasionally attempt to ridicule this argument by pretending to interpret it as meaning that the difference between x and x_0 can be small at the same time as that between x_0 and $\hat{x}$ is large.[16] As we have stated above, we are not drawing a distinction between these two cases (there is none) but between conditioning on the hypotheses 'whatever x_0 may be' and 'whatever $\hat{x}$ may be'. Tracing this back to Bayes's theorem, what goes wrong is that those who do not wish to use it in a legitimate way – on account of certain scruples – have no scruples at all about using it in a manifestly illegitimate way. That is to say, they ignore one of the factors (the prior probability) altogether and treat the other (the likelihood) as though it in fact meant something other than it actually does. This is the same mistake as is made by someone who has scruples about measuring the arms of a balance (having only a tape-measure at his disposal, rather than a high precision instrument) but is willing to assert that the heavier load will always tilt the balance (thereby implicitly assuming, although without admitting it, that the arms are of equal length!).

These same comments apply, essentially unaltered, to the case of hypothesis testing and to other topics (and so we shall not bother to repeat them), because they relate to the essence of the whole 'objectivistic' approach to statistics. One important consequence is the realization that objectivistic forms of significance test do not obey the likelihood principle. These are tests in which, for example, one rejects the null

[15] If we call X the true value, Y the estimated value and $Z = Y - X$ the error, it is clear that the distribution of Z given $X = x_0$ is not the same thing as the distribution of Z given $Y = y_0$. If $f(x, y)$ represents the joint density for (X, Y), then, in the two cases, the distributions of Z are given by $Kf(x_0, x_0 + z)$ and $Kf(y_0 - z, y_0)$, respectively. These can only coincide if X and Y both have improper uniform densities and Z is independent (i.e. $f(x,y) = Kg(y - x)$ with $K = 0$, in the usual sense). In the case of Section 12.3.2 (X and Z independent and normally distributed), $f(z|x)$ is independent of x by hypothesis (normal distribution $N(0, \sigma/\sqrt{n})$ but $f(z|y)$, as is shown in (16), although still normal and having the same variance, has nonzero prevision:

$$\mathbf{P}(Z \mid Y = y_0) = m_f - y_0 = (m_0 - y_0)/\left[1 + n\sigma^{-2}/\sigma_0^{-2}\right],$$

where y_0 was denoted in equation 12.16 by $n\bar{x}$. The term $m_f - y_o$ only vanishes if we take $\sigma_0 = \infty$; i.e. the improper uniform distribution over ± ∞.

Objectivists will probably argue that, as a rule, really large errors do not occur and that if they do one notices the fact and rejects the observation. However, the rejection of a complete, coherent formulation cannot be justified under the pretext that if something does not seem to work one can always get out of trouble by resorting to expedients which themselves cannot be justified (neither in the new, patched-up formulation, nor in the coherent one).

[16] This is a rather imprecise objection, open to several interpretations. Only laymen (so far as this topic is concerned, anyway) could take it literally as providing evidence of an oversight. That we are dealing with two different things (see the explanation in the text) is clear, not only to the Bayesian but also to objectivists of the Neyman–Pearson school. The difference is that the latter deliberately choose to base themselves on considerations of the form 'whatever x_0 may be', in order to avoid the Bayesian formulation (assuming arguments based upon the former considerations to be valid, despite the fact that they do not have the same meaning as those in the Bayesian framework, and, indeed going so far as to claim the former as 'modern', and the latter as 'old fashioned'). R.A. Fisher, on the other hand, attempted to create a fusion of the two. It seems to me that he felt the need for the Bayesian form of conclusion (although he expressed it in an illusory manner by means of an undefinable 'fiducial probability'), but wanted to approach the problem from the opposite direction (an approach rather like that of Neyman).

hypothesis $\theta = \theta_0$ because some given function $t(x)$ of the observed data (a statistic) has 'too large' a value (lying outside some given confidence interval; i.e. in the 'tails' of the distribution of t). The point concerning the likelihood principle is clear, because, for the objectivist, the confidence interval is one in which, with $100\beta\%$ probability, $t(x)$ must lie given θ_0 (and not vice versa!). An example of this is given in Lindley, Vol. II, pp. 68–69.

These strictures do not imply, however, that the conclusions cannot, in practice, be satisfactory for most applications. Referring to our example, it will, in fact, be very rare for x_0 not to be close to $\hat{x}$. However, why should we blind ourselves to the possibility of it being otherwise? Why should we stick to the standard conclusion even in cases where we are suspicious? Why should we be forced to ignore facts which, if we do not wish to shut our eyes to them, should lead us to be suspicious?

In any case, a Bayesian analysis will indicate within what limits, and under what conditions, any particular method is approximately valid and what needs to be done (following Lindley's example, perhaps) in order to turn it into an exact and acceptable procedure.

12.6.5. The method of *maximum likelihood* was developed in particular by R.A. Fishe, and, although it was known previously, it was through his work that it came to prominence.

In its crudest form, as a method of point estimation, it consists in taking the estimate of a parameter θ as the value (or vector etc.) $\hat{\theta}$ that gives the (absolute) maximum of the likelihood for θ given by the observations x. One can give this a Bayesian interpretation as the estimate of θ given by the *mode* of the posterior distribution, assuming the prior to be uniform (since the posterior *coincides*[17] with the likelihood in this case, the point maximizing the former also maximizes the latter).

The most useful application of the concept is in providing a *normal* approximation to the posterior distribution (which can then, if one wishes, be used to give an interval estimate).

If we consider the standard case of exchangeability, that is repeated observations with the same density $f(x|\theta)$, the likelihood is the product, as we have seen many times before, and its logarithm is given by

$$L_n(x \mid \theta) = L(x \mid \theta) = \log p(x \mid \theta) = \sum_{h=1}^{n} \log f(x_h \mid \theta). \tag{12.31}$$

The logarithm is used simply for convenience and the function L, to use the standard notation, is called the log-likelihood (the subscript n usually being omitted).

As n increases, the influence of the prior distribution π_0 on the posterior π_n becomes smaller and smaller, the fixed factor being overwhelmed by the n factors of the form $f(x_h|\theta)$. This was clear even in Section 12.2.5 (see equation 12.3′) and especially so in the examples we studied (see equation 12.15 of Section 12.3.2 etc.). This means that the more observations that are available, the more their influence is predominant in

[17] Recall that it is not really correct to say 'coincides', because the likelihood is a point function, whereas the density depends on the point *and* on the measure (see the final remark of Section 12.3.4).

determining our posterior opinions and the less significant the prior opinion becomes. This is what we would expect.

Because of this (a fact which, incidentally, has been appreciated for quite a while, and was well illustrated by Poincaré), the difficulties we mentioned relating to the evaluation of the prior probabilities turn out to be less serious from a practical point of view. We are not saying that the problem disappears, but that it becomes possible to deal with it in a satisfactory manner by making precise the conditions and the limits within which it is possible to replace a given prior distribution by the uniform, for example, without causing any serious distortion.

In general, and under fairly weak conditions, the likelihood, for large n, is sharply peaked around its maximum, so that the maximum likelihood estimate $\hat{\theta}$ is, as it stands, quite informative (and this is true, in particular, in the case of the normal distribution). A point estimate on its own, however, is never very satisfactory and it is fortunate that the maximum likelihood approach enables us to improve on this by also providing the variance, not of $\pi_n(\theta)$ itself, but of the normal approximation to it in a neighbourhood of $\hat{\theta}$. In fact, we have

$$\sigma_n^{-2} = -\frac{\partial^2}{\partial\theta^2} L_n\left(x \mid \hat{\theta}\right).^{18} \tag{12.32}$$

A rough argument will suffice to show why this is so:[19] expanding $L_n(\theta) = L(x|\theta)$ around $\theta \simeq \hat{\theta}$, we obtain

$$L_n\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right) L_n'\left(\hat{\theta}\right) + \frac{1}{2}\left(\theta - \hat{\theta}\right)^2 L_n''\left(\hat{\theta}\right) + \ldots,$$

and the density (again, in a neighbourhood of $\hat{\theta}$) is given by

$$\pi_n\left(\theta\right) = Kss\pi_0\left(\theta\right) \mathrm{e}^{L_n(x|\theta)} \simeq K.\mathrm{e}^{+\frac{1}{2}\left(\theta-\hat{\theta}\right)^2 L_n''\left(\hat{\theta}\right)}, \tag{12.33}$$

because: (i) $\pi_0(\theta)$ (in the small neighbourhood of $\hat{\theta}$ in which $L_n(\theta)$ is large) is practically constant (and equal to $\pi_0(\hat{\theta})$); (ii) $L_n(\hat{\theta})$, which is constant, can be subsumed in K; (iii) $L_n'(\hat{\theta}) = 0$, because L_n is maximized at $\hat{\theta}$; and (iv) we can neglect terms beyond those of second order.

The term $L''(\hat{\theta})$ is called the 'information' (but must not be confused with the concept as used in information theory; see Chapter 3, 3.8.5). The result can be extended to the case where θ is a vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$. The distribution is then multivariate normal and a natural generalization of equation 12.32 defines the *information matrix* as the inverse of the variance-covariance matrix:

$$I_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} L\left(x \mid \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s\right). \tag{12.34}$$

[18] Here, and in equation 12.34, the derivative of $L_n(x|\theta)$ is evaluated at $\theta = \hat{\theta}$.

[19] This is basically the same argument as that given in Chapter 11, 11.4.4.

In other words, the I_{ij} give the coefficients of the $(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$ terms in the quadratic form – Q appearing in the density $K.\mathrm{e}^{-\frac{1}{2}Q}$.

12.6.6. Although our account has been very brief, it has dealt with several of the most important topics of mathematical statistics, both from the Bayesian and the objectivistic standpoints. Moreover, we have indicated the main points of departure of the two approaches.

In particular, we note that, in practical terms, the situation is altered by the fact of whether n is large (when we enter the realm of so-called *large-sample* theory) or small (*small-sample* theory). In the first case, practically any method works; whereas, in the second, different methods lead, in general, to very different conclusions. The Bayesian approach is part of a coherent, formal theory, which rules out any conceptual obscurity. On the other hand, in the case of small samples the conclusions are strongly dependent on prior opinions, which may vary greatly from one individual to another. This is a genuine unavoidable fact, but it is not a drawback of the Bayesian approach. It would be a drawback if it were an unnecessary complication but the fact is that if complications do actually exist the drawbacks and errors stem from ignoring them and providing pie-in-the-sky solutions that do not take them into account (as in the objectivistic approach).

Anyway, in concluding this summary I should like to quote the following words of Lindley (Vol. II, Preface, p. xii):[20]

> 'Most of modern statistics (*i.e. that of the objectivistic school*) is perfectly sound in practice; it is done for the wrong reason. Intuition has saved the statistician from error. My contention is that the Bayesian method justifies what he has been doing (*by reinterpreting and correcting it*) and develops new methods that the "orthodox" approach lacks.'

12.7 The Connections with Decision Theory

12.7.1. It is not our intention to discuss this topic at all thoroughly, nor would it be possible to do so within the limits of the present outline treatment. Had we wished to do so, however, we could have set the whole of this chapter within a decision-theoretic framework. What we shall do is to clarify some of the areas in which decision theory offers additional insights into certain of the problems of mathematical statistics and into the comparison between the Bayesian and the more fashionable objectivistic approaches.

We mentioned this topic briefly at the end of Chapter 3 (and here and there in the sequel), where we observed that coherence required us to adopt the criterion of maximizing (expected) utility as the basis of decision making.

Basically, it tells us that we should arrive at a decision by first considering the individual increments of utility attached to the consequences of the various possible decisions, and then weighting these by the respective probabilities. A decision must, therefore, be based on probabilities; that is the posterior probabilities as evaluated on the basis of all

[20] The explanations in parentheses, together with the quotation marks for 'orthodox', are not part of the original.

information so far available. This is the main point to note. In order to make decisions, we first require a statistical theory that provides conclusions in the form of posterior probabilities. The Bayesian approach does this; other approaches explicitly refuse to do this.

Indeed, objectivistic approaches to statistics bend over backwards to give nonprobabilistic answers to probabilistic questions, expressing them in YES–NO terms, as in the logic of certainty. More specifically, they talk in terms of 'accepting' or 'rejecting' a given hypothesis on the basis of some given test and, although some hesitate to go this far, occasionally one hears that 'to accept an hypothesis' means 'to agree to behave as if it were certainly true'. This is nonsense. One should not behave 'as if an hypothesis were certain' unless it actually is regarded as certain. If it is not, then we cannot decide how to behave until we have attributed to it some probability p. The appropriate behaviour is then that which, on the basis of p, is calculated to maximize expected utility.

12.7.2. Some authors (notably R.A. Fisher) criticize the application of these ideas to problems of scientific inference, regarding them as essentially economic in nature and incompatible with pure research. We could object that even in the scientific field one cannot escape having to weigh up favourable and unfavourable consequences, but a more decisive reply stems from the fact that these 'economic' arguments reveal the necessity of making sure that opinions cohere. In particular, they show that one must pass from prior to posterior opinions in conformity with Bayes's theorem, and that this is the case no matter whether we are contemplating a bet, a business decision, or simply recording our conclusion for use in a scientific context.

There do not exist two entirely different forms of valid reasoning, one suitable in a commercial context, the other for pure research. No one working in the scientific field considers it beneath him to use the same arithmetic operations, or calculating machines, as are needed for commercial purposes. There is only one theory and it does not matter whether it is used for utilitarian purposes, or for pure research, or simply studied for its own sake.

12.7.3. It is interesting to note that the movement towards a decision-theoretic point of view began within the framework of objectivistic theory. Above all, this was the result of Abraham Wald's introduction of the idea of associating a loss with an incorrect decision, taking, as an example of this, the acceptance of an hypothesis i given that hypothesis j is true (loss $= L_{ij}$, zero if $i = j$). However, this does not entirely remove the unsatisfactory identification of the decision as the 'acceptance of an hypothesis'. The necessary step involves singling out the individual possible 'actions' – choice among which corresponds to a probabilistic assessment – rather than acceptance of the various hypotheses. Some criteria of decision making are taken over from other contexts, without examining closely their suitability for the problem under consideration (for example, the minimax criterion is considered acceptable, even though it corresponds to a different situation, that of competitive uncertainty – i.e. games theory). However, Wald's formulation did result in the explicit reintroduction of prior probabilities and hence Bayesian theory (albeit in a formal sense, without involving the subjectivistic interpretation).

Other movements in this direction have sprung from criticisms of various paradoxes and defects within the objectivistic framework itself. In order to remove these, it became clear that a Bayesian formulation was required. In this context, the contributions of I.J. Good, D.V. Lindley and L.J. Savage deserve explicit mention.

In addition there has been a great deal of research into the economics of uncertainty. Through the work of von Neumann and Morgenstern this gave new life and impetus to the study of utility theory, which had long been neglected (although various scholars – Daniel Bernoulli in the past, and F.P. Ramsey more recently – had shown an interest in it). These various strands of research found their culmination in the work of L.J. Savage, *The Foundations of Statistics* (1954) (so far, that is, as the theme of this book is concerned; the revision of statistical methodology from a Bayesian point of view is more recent, and is still continuing).

12.7.4. Finally, we should note that decision theory has very important things to say about questions relating to the planning of experiments for statistical purposes. In other words, planning experiments in order to improve the information on the basis of which decisions are to be taken.

One aspect of this involves the techniques of such experiments, these being studied in order to optimize the outcome; that is to obtain the most valid and useful information at the least possible cost. It would take us too long even to just mention the most important problems and methods, which have been extensively studied in the literature. It will suffice to simply point out that a vast amount of research has been done and that its enormous contribution to technological progress cannot be properly appreciated unless one examines a number of examples.

So far as we are concerned, it is the more basic aspect of all this which interests us. We are referring to the fact that the reasons which make clear to us the correct form of argument for reaching useful, practical decisions, and that required for reaching conceptually valid conclusions, are the *same* in both cases.

In fact, the seemingly 'new' problem, '*what information is it most useful to obtain before making the decision*?', can be considered as it stands within our previous formulation, underlining yet again its general and comprehensive nature. It suffices to include as possible 'actions' not only those of the original formulation – that is those relating to the 'final decision' – but also the various possible choices of experimental procedure and model-building which lead up to it. The value of any piece of information (in the context of a particular decision problem) can be measured as the increment of expected utility deriving from it (or, in the simplest case, the increment of expected gain). This value is always positive (although it could be zero; if the worst comes to worst, we can always take the decision without taking into account the additional information) but there is usually some cost incurred (for labour, time etc.). The net gain from the information (or, more precisely, from the decision to obtain it) is the difference between its value and its cost. The optimal decision (regarding what information one should seek to obtain) is given by that for which the difference between the value and the cost is maximized. In general, the process of collecting information may be quite complicated (performed sequentially, in a number of stages, with a built-in arrangement for subsequent choices to depend on the information obtained initially, and so on). From a conceptual point of view, our general approach can cope with all this without requiring any modification.

In this context, it is clear from a practical point of view that there is a need for coherence not only for each individual decision but also at an overall level, linking the individual steps together. Such a requirement is perfectly obvious if problems are set out in a detailed fashion within their natural probabilistic setting, but it tends to be overlooked if one gets used to dealing with problems on a fragmentary basis.

Typical of the confusion that can arise is the statement that the 'minimax' procedure (in decision theory) is coherent. In actual fact, it is coherent for each individual application, because it turns out to be *Bayesian* under the choice of a particular prior distribution (and any one is free to choose it if they wish). It corresponds to the choice of the *least favourable* distribution, one that would be used by an opponent who wished to make things as difficult as possible for us. This analogy with games theory – more precisely, with two-person zero-sum games, that is those in which one person's loss is the other's gain – is often emphasized by referring to a statistical decision as a 'game against Nature'. The analogy only goes through, however, if one assumes 'Nature to be malevolent'.

Apart from any reservations one might have about this latter hypothesis,[21] we see at once that it cannot be applied in every case. In fact, if we simultaneously consider a number of decisions all depending on the same event, this approach will certainly lead to contradictions, because the least favourable distribution for one decision will not, in general, be the least favourable for the others. Nature (nor any other opponent for that matter) cannot be so evil-minded as to simultaneously adopt distributions – or '*strategies*' as they are called in games theory – which necessarily put us in a least favourable position for *any individual* decision problem that we might wish to consider.[22] As an obvious analogy, anyone being pursued by a number of hunters coming at him from different directions cannot escape in the opposite direction to all of them.

12.7.5. It might appear that these, our final considerations, have only been made possible by the long and wearisome journey that has gone before. In fact, this is not so. If one sticks to the approach that we have advocated throughout, all this – and let us repeat it once more, so that there is no doubt – is obvious. The time and energy was required for the long excursion that we made into objectivistic territory – a necessary journey, undertaken not as an end in itself, but in order to dispel the notion that an objectivistic formulation could constitute an acceptable, alternative approach. That journey is now over and our work is done. Free at last from paradoxes and contradictions, we emerge from our sea of troubles.

[21] Some people attempt to justify it as a 'conservative policy' for anyone wishing 'to guard themselves against the risk of the worst happening to them'. The solution to this, if there is one, lies in choosing a very convex utility function, not in deliberately distorting one's opinions; this can only result in a worse decision, and is therefore unacceptable.

[22] Anyone wishing to take seriously the hypothesis that Nature is ill-disposed towards him, should adopt a prior distribution that is least favourable *over the whole range of decisions* confronting him, and involving the circumstances under consideration. This would involve applying the minimax criterion to the single, compound problem (taking the entire complex of possible decisions as a single decision), or, alternatively (if the cases are independent), solving each individual decision one at a time, but in terms of what 'Nature's evil-minded strategy' would be over the whole complex of decisions (a very different situation from individual applications of minimax). This is the same as the distinction between minimizing a sum of functions $f(x) = \sum_n f_n(x)$ – i.e. $f(\xi)$ at the value ξ where the sum obtains its minimum – and evaluating $\sum_n f_n(\xi_n)$ the sum at the individual minima (as if x could assume simultaneously – perhaps being evil-minded – the different values $x = \xi_n$).