Chapter 2

Foundations

Summary

The concept of rationality is explored in the context of representing beliefs or choosing actions in situations of uncertainty. An axiomatic basis, with intuitive operational appeal, is introduced for the foundations of decision theory. The dual concepts of probability and utility are formally defined and analysed within this context. The criterion of maximising expected utility is shown to be the only decision criterion which is compatible with the axiom system. The analysis of sequential decision problems is shown to reduce to successive applications of the methodology introduced. Statistical inference is viewed as a particular decision problem which may be analysed within the framework of decision theory. The logarithmic score is established as the natural utility function to describe the preferences of an individual faced with a pure inference problem. Within this framework, the concept of discrepancy between probability distributions and the quantification of the amount of information in new data are naturally defined in terms of expected loss and expected increase in utility, respectively.

2.1 BELIEFS AND ACTIONS

We spend a considerable proportion of our lives, both private and professional, in a state of uncertainty. This uncertainty may relate to past situations, where direct knowledge or evidence is not available, or has been lost or forgotten; or to present and future developments which are not yet completed. Whatever the circumstances, there is a sense in which all states of uncertainty may be described in the same way: namely, an individual feeling of incomplete knowledge in relation to a specified situation (a feeling which may, of course, be shared by other individuals). And yet it is obvious that we do not attempt to treat all our individual uncertainties with the same degree of interest or seriousness.

Many feelings of uncertainty are rather insubstantial and we neither seek to analyse them, nor to order our thoughts and opinions in any kind of responsible way. This typically happens when we feel no actual or practical involvement with the situation in question. In other words, when we feel that we have no (or only negligible) capacity to influence matters, or that the possible outcomes have no (or only negligible) consequences so far as we are concerned. In such cases, we are not motivated to think carefully about our uncertainty either because nothing depends on it, or the potential effects are trivial in comparison with the effort involved in carrying out a conscious analysis.

On the other hand, we all regularly encounter uncertain situations in which we at least aspire to behave "rationally" in some sense. This might be because we face the direct practical problem of choosing from among a set of possible actions, where each involves a range of uncertain consequences and we are concerned to avoid making an "illogical" choice. Alternatively, we might be called upon to summarise our beliefs about the uncertain aspects of the situation, bearing in mind that others may subsequently use this summary as the basis for choosing an action. In this case, we are concerned that our summary be in a form which will enable a "rational" choice to be made at some future time. More specifically, we might regard the summary itself, i.e., the choice of a particular mode of representing and communicating our beliefs, as being a form of action to which certain criteria of "rationality" might be directly applied.

Our basic concern in this chapter is with exploring the concept of "rationality" in the context of representing beliefs or choosing actions in situations of uncertainty. To choose the best among a set of actions would, in principle, be immediate if we had perfect information about the consequences to which they would lead. So far as this work is concerned, interesting decision problems are those for which such perfect information is not available, and we must take *uncertainty* into account as a major feature of the problem.

It might be argued that there are complex situations where we *do* have complete information and yet still find it difficult to take the best decision. Here, however, the difficulty is *technical*, not conceptual. For example, even though we have, in principle, complete information, it is typically not easy to decide what is the optimal strategy to rebuild a Rubik cube or which is the cheapest diet fulfilling specified nutritional requirements. We take the view that such problems are purely technical. In the first case, they result from the large number of possible strategies;

in the second, they reduce to the mathematical problem of finding a minimum under certain constraints. But in neither case is there any doubt about the decision criterion to be used. In this work we shall not consider these kinds of combinatorial or mathematical programming problems, and we shall assume that in the presence of complete information we can, in principle, always choose the best alternative.

Our concern, instead, is with the *logical process of decision making in situations of uncertainty*. In other words, with the decision criterion to be adopted when we do not have complete information and are thus faced with, at least some, elements of uncertainty.

To avoid any possible confusion, we should emphasise that we do not interpret "actions in situations of uncertainty" in a narrow, directly "economic" sense. For example, within our purview we include the situation of an individual scientist summarising his or her own current beliefs following the results of an experiment; or trying to facilitate the task of others seeking to decide upon their beliefs in the light of the experimental results.

It is assumed in our approach to such problems that the notion of "rational belief" cannot be considered separately from the notion of "rational action". Either a statement of beliefs in the light of available information is, actually or potentially, an input into the process of choosing some practical course of action,

> . . . it is not asserted that a belief . . . does actually lead to action, but would lead to action in suitable circumstances; just as a lump of arsenic is called poisonous not because it actually has killed or will kill anyone, but because it would kill anyone if he ate it (Ramsey, 1926).

or, alternatively, a statement of beliefs might be regarded as an end in itself, in which case the choice of the form of statement to be made constitutes an action,

> Frequently, it is a question of providing a convenient summary of the data . . . In such cases, the emphasis is on the inference rather than the decision aspect of problem, although formally it can still be considered a decision problem if the inferential statement itself is interpreted as the decision to be taken (Lehmann, 1959/1986).

We can therefore explore the notion of "rationality" for both beliefs and actions by concentrating on the latter and asking ourselves what kinds of rules should govern preference patterns among sets of alternative actions in order that choices made in accordance with such rules commend themselves to us as "rational", in that they cannot lead us into forms of behavioural inconsistency which we specifically wish to avoid.

In Section 2.2, we describe the general structure of problems involving choices under uncertainty and introduce the idea of preferences between options. In Section 2.3, we make precise the notion of "rational" preferences in the form of axioms.

We describe these as *principles of quantitative coherence* because they specify the ways in which preferences need to be made *quantitatively* precise and fit together, or *cohere*, if "illogical" forms of behaviour are to be avoided. In Sections 2.4 and 2.5, we prove that, in order to conform with the principles of quantitative coherence, degrees of belief about uncertain events should be described in terms of a (finitely additive) *probability measure*, relative values of individual possible consequences should be described in terms of a *utility function*, and the rational choice of an action is to select one which has the *maximum expected utility*.

In Section 2.6, we discuss sequential decision problems and show that their analysis reduces to successive applications of the maximum expected utility methodology; in particular, we identify the design of experiments as a particular case of a sequential decision problem. In Section 2.7, we make precise the sense in which choosing a form of a statement of beliefs can be viewed as a special case of a decision problem. This identification of inference as decision provides the fundamental justification for beginning our development of Bayesian Statistics with the discussion of decision theory. Finally, a general review of ideas and references is given in Section 2.8.

2.2 DECISION PROBLEMS

2.2.1 Basic Elements

We shall describe any situation in which choices are to be made among alternative courses of action with uncertain consequences as a *decision problem*, whose structure is determined by three basic elements:

(i) a set $\{a_i,\, i \in I\}$ of available *actions*, one of which is to be selected;

(ii) for each action a_i, a set $\{E_j,\, j \in J\}$ of *uncertain events*, describing the uncertain outcomes of taking action a_i;

(iii) corresponding to each set $\{E_j,\, j \in J\}$, a set of *consequences* $\{c_j,\, j \in J\}$.

The idea is as follows. Suppose we choose action a_i; then one and only one of the uncertain events $E_j,\, j \in J$, occurs and leads to the corresponding consequence $c_j,\, j \in J$. Each set of events $\{E_j,\, j \in J\}$ forms a *partition* (an exclusive and exhaustive decomposition) of the total set of possibilities. Naturally, both the set of consequences and the partition which labels them may depend on the particular action considered, so that a more precise notation would be $\{E_{ij},\, j \in J_i\}$ and $\{c_{ij},\, j \in J_i\}$ for each action a_i. However, to simplify notation, we shall omit this dependence, while remarking that it should always be borne in mind. We shall come back to this point in Section 2.6.

In practical problems, the labelling sets, I and J (for each i), are typically finite. In such cases, the decision problem can be represented schematically by means of a decision tree as shown in Figure 2.1.

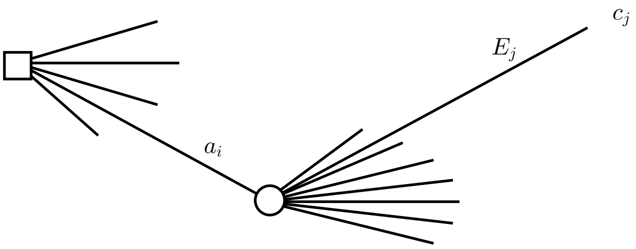

Figure 2.1 *Decision tree*

The square represents a decision node, where the choice of an action is required. The circle represents an uncertainty node where the outcome is beyond our control. Following the choice of an action and the occurrence of a particular event, the branch leads us to the corresponding consequence.

Of course, most practical problems involve sequential considerations but, as shown in Section 2.6, these reduce, essentially, to repeated analyses based on the above structure.

It is clear, either from our general discussion, or from the decision tree representation, that we can formally identify any a_i, $i \in I$, with the combination of $\{E_j,\ j \in J\}$ and $\{c_j,\ j \in J\}$ to which it leads. In other words, to choose a_i is to opt for the uncertain scenario labelled by the pairs (E_j, c_j), $j \in J$. We shall write $a_i = \{c_j \mid E_j,\ j \in J\}$ to denote this identification, where the notation $c_j \mid E_j$ signifies that event E_j leads to consequence c_j, i.e., that $a_i(E_j) = c_j$.

An individual's perception of the state of uncertainty resulting from the choice of any particular a_i is very much dependent on the *information currently available*. In particular, $\{E_j,\ j \in J\}$ forms a partition of the total set of relevant possibilities as the individual decision-maker now perceives them to be. Further information, of a kind which leads to a restriction on what can be regarded as the total set of possibilities, will change the perception of the uncertainties, in that some of the E_j's may become very implausible (or even logically impossible) in the light of the new information, whereas others may become more plausible. It is therefore of considerable importance to bear in mind that a representation such as Figure 2.1 only captures the structure of a decision problem as perceived at a particular point in time. Preferences about the uncertain scenarios resulting from the choices of actions depend on *attitudes to the consequences involved and assessments of the uncertainties attached to the corresponding events*. The latter are clearly subject to change as new information is acquired and this may well change overall preferences among the various courses of action.

The notion of *preference* is, of course, very familiar in the everyday context of actual or potential choice. Indeed, an individual decision-maker often prefaces an actual choice (from a menu, an investment portfolio, a range of possible forms of medical treatment, a textbook of statistical methods, etc.) with the phrase "I prefer… " (caviar, equities, surgery, Bayesian procedures, etc.). To prefer action a_1 to action a_2 means that if these were the only two options available a_1 would be chosen (conditional, of course, on the information available at the time). In everyday terms, the idea of indifference between two courses of action also has a clear operational meaning. It signifies a willingness to accept an externally determined choice (for example, letting a disinterested third party choose, or tossing a coin).

In addition to *representing the structure* of a decision problem using the three elements discussed above, we must also be able to *represent the idea of preference* as applied to the comparison of some or all of the pairs of available options. We shall therefore need to consider a fourth basic element of a decision problem:

(iv) the relation $\leq$, which expresses the individual decision-maker's preferences between pairs of available actions, so that $a_1 \leq a_2$ signifies that a_1 *is not preferred to* a_2.

These four basic elements have been introduced in a rather informal manner. In order to study decision problems in a precise way, we shall need to reformulate these concepts in a more formal framework. The development which follows, here and in Section 3.3, is largely based on Bernardo, Ferrándiz and Smith (1985).

2.2.2 Formal Representation

When considering a particular, concrete decision problem, we do not usually confine our thoughts to *only* those outcomes and options explicitly required for the specification of that problem. Typically, we *expand our horizons* to encompass analogous problems, which we hope will aid us in ordering our thoughts by providing suggestive points of reference or comparison. The collection of uncertain scenarios defined by the original concrete problem is therefore implicitly embedded in a somewhat wider framework of actual and hypothetical scenarios. We begin by describing this *wider frame of discourse* within which the comparisons of scenarios are to be carried out. It is to be understood that the initial specification of any such particular frame of discourse, together with the preferences among options within it, are dependent on the decision-maker's overall state of information at that time. Throughout, we shall denote this initial state of mind by M_0.

We now give a formal definition of a decision problem. This will be presented in a rather compact form; detailed elaboration is provided in the remarks following the definition.

Definition 2.1. (***Decision problem***). *A decision problem is defined by the elements* $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq)$*, where:*

(i) $\mathcal{E}$ *is an algebra of relevant events,* E_j*;*

(ii) *$\mathcal{C}$ is a set of possible consequences, c_j;*

(iii) *$\mathcal{A}$ is a set of options, or potential acts, consisting of functions which map finite partitions of Ω, the certain event in $\mathcal{E}$, to compatibly-dimensioned, ordered sets of elements of $\mathcal{C}$;*

(iv) *$\leq$ is a preference order, taking the form of a binary relation between some of the elements of $\mathcal{A}$.*

We now discuss each of these elements in detail. Within this wider frame of discourse, an individual decision-maker will wish to consider the uncertain events judged to be *relevant* in the light of the initial state of information M_0. However, it is natural to assume that if $E_1 \in \mathcal{E}$ and $E_2 \in \mathcal{E}$ are judged to be relevant events then it may also be of interest to know about their joint occurrence, or whether at least one of them occurs. This means that $E_1 \cap E_2$ and $E_1 \cup E_2$ should also be assumed to belong to $\mathcal{E}$. Repetition of this argument suggests that $\mathcal{E}$ should be closed under the operations of arbitrary finite intersections and unions. Similarly, it is natural to require $\mathcal{E}$ to be closed under complementation, so that $E^c \in \mathcal{E}$. In particular, these requirements ensure that the *certain event* Ω and the *impossible event* $\emptyset$, both belong to $\mathcal{E}$. Technically, we are assuming that the class of relevant events has the structure of an *algebra*. (However, it can certainly be argued that this is too rigid an assumption. We shall provide further discussion of this and related issues in Section 2.8.4.)

As we mentioned when introducing the idea of a wider frame of discourse, the algebra $\mathcal{E}$ will consist of what we might call the *real-world* events (that is, those occurring in the structure of any concrete, actual decision problem that we may wish to consider), together with any other *hypothetical* events, which it may be convenient to bring to mind as an aid to thought. The class $\mathcal{E}$ will simply be referred to as the *algebra of (relevant) events*.

We denote by $\mathcal{C}$ the set of all consequences that the decision-maker wishes to take into account; preferences among such consequences will later be assumed to be independent of the state of information concerning relevant events. The class $\mathcal{C}$ will simply be referred to as the *set of (possible) consequences*.

In our introductory discussion we used the term *action* to refer to each potential act available as a choice at a decision node. Within the wider frame of discourse, we prefer the term *option*, since the general, formal framework may include hypothetical scenarios (possibly rather far removed from potential concrete actions).

So far as the definition of an option as a function is concerned, we note that this is a rather natural way to view options from a mathematical point of view: an option consists precisely of the linking of a partition of Ω, $\{E_j,\ j \in J\}$, with a corresponding set of consequences, $\{c_j,\ j \in J\}$. To represent such a mapping we shall adopt the notation $\{c_j \mid E_j,\ j \in J\}$, with the interpretation that event E_j leads to consequence c_j, $j \in J$.

It follows immediately from the definition of an option that the ordering of the labels within J is irrelevant, so that, for example, the options $\{c_1 \,|\, E,\ c_2 \,|\, E^c\}$, and $\{c_2 \,|\, E^c,\ c_1 \,|\, E\}$ are identical, and forms such as $\{c \,|\, E_1,\ c \,|\, E_2,\ c_j \,|\, E_j, j \in J\}$ and $\{c \,|\, E_1 \cup E_2,\ c_j \,|\, E_j,\ j \in J\}$ are completely equivalent. Which form is used in any particular context is purely a matter of convenience. Sometimes, the interpretation of an option with a rather cumbersome description is clarified by an appropriate reformulation. For example, $a = \{c_1 \,|\, E \cap G,\ c_2 \,|\, E^c \cap G,\ c_3 \,|\, G^c\}$ may be more compactly written as $a = \{a_1 \,|\, G, c_3 \,|\, G^c\}$, with $a_1 = \{c_1 \,|\, E, c_2 \,|\, E^c\}$. Thus, if

$$a = \{c_{k(j)} \,|\, E_{k(j)} \cap F_j, k(j) \in K_j, j \in J\}, \quad a_j = \{c_{k(j)} \,|\, E_{k(j)}, k(j) \in K_j\},$$

we shall use the *composite function notation* $a = \{a_j \,|\, F_j, j \in J\}$. In all cases, the ordering of the labels is irrelevant. The class $\mathcal{A}$ of options, or potential actions, will simply be referred to as the *action space*.

In defining options, the assumption of a *finite* partition into events of $\mathcal{E}$ seems to us to correspond most closely to the structure of practical problems. However, an extension to admit the possibility of *infinite* partitions has certain mathematical advantages and will be fully discussed, together with other mathematical extensions, in Chapter 3.

In introducing the preference binary relation $\leq$, we are not assuming that all pairs of options $(a_1, a_2) \in \mathcal{A} \times \mathcal{A}$ can necessarily be related by $\leq$. If the relation *can* be applied, in the sense that either $a_1 \leq a_2$ or $a_2 \leq a_1$ (or both), we say that a_1 is not preferred to a_2, or a_2 is not preferred to a_1 (or both). From $\leq$, we can derive a number of other useful binary relations.

Definition 2.2. (***Induced binary relations***).

(i) $a_1 \sim a_2 \iff a_1 \leq a_2$ *and* $a_2 \leq a_1$.

(ii) $a_1 < a_2 \iff a_1 \leq a_2$ *and it is not true that* $a_2 \leq a_1$.

(iii) $a_1 \geq a_2 \iff a_2 \leq a_1$.

(iv) $a_1 > a_2 \iff a_2 < a_1$.

Definition 2.2 is to be understood as referring to any options a_1, a_2 in $\mathcal{A}$. To simplify the presentation we shall omit such universal quantifiers when there is no danger of confusion. The induced binary relations are to be interpreted to mean that a_1 is equivalent to a_2 if and only if $a_1 \sim a_2$, and a_1 is strictly preferred to a_2 if and only if $a_1 > a_2$. Together with the interpretation of $\leq$, these suffice to describe all cases where pairs of options can be compared.

We can identify individual consequences as special cases of options by writing $c = \{c \,|\, \Omega\}$, for any $c \in \mathcal{C}$. Without introducing further notation, we shall simply regard c as denoting either an element of $\mathcal{C}$, or the element $\{c \,|\, \Omega\}$ of $\mathcal{A}$. There will be no danger of any confusion arising from this identification. Thus, we shall write $c_1 \leq c_2$ if and only if $\{c_1 \,|\, \Omega\} \leq \{c_2 \,|\, \Omega\}$ and say that consequence c_1 is not preferred to consequence c_2. Strictly speaking, we should introduce a new symbol to replace $\leq$ when referring to a preference relation over $\mathcal{C} \times \mathcal{C}$, since $\leq$ is defined over $\mathcal{A} \times \mathcal{A}$. In fact, this parsimonious abuse of notation creates no danger of confusion and we shall routinely adopt such usage in order to avoid a proliferation of symbols. We shall proceed similarly with the binary relations $\sim$ and $<$ introduced in Definition 2.2. To avoid triviality, we shall later formally assume that there exist at least two consequences c_1 and c_2 such that $c_1 < c_2$.

The basic preference relation between options, $\leq$, conditional on the initial state of information M_0, can also be used to define a binary relation on $\mathcal{E} \times \mathcal{E}$, the collection of all pairs of *relevant events*. This binary relation will capture the intuitive notion of one event being "more likely" than another. Since, once again, there is no danger of confusion, we shall further economise on notation and also use the symbol $\leq$ to denote this new uncertainty binary relation between events.

Definition 2.3. (***Uncertainty relation***).

$$E \leq F \iff \textit{for all } c_1 < c_2,\ \{c_2 \,|\, E, c_1 \,|\, E^c\} \leq \{c_2 \,|\, F, c_1 \,|\, F^c\};$$

*we then say that E is **not more likely** than F.*

The intuitive content of the definition is clear. If we compare two dichotomised options, involving the same pair of consequences and differing only in terms of their uncertain events, we will prefer the option under which we feel it is "more likely" that the preferred consequence will obtain. Clearly, the force of this argument applies independently of the choice of the particular consequences c_1 and c_2, provided that our preferences between the latter are assumed independent of any considerations regarding the events E and F.

Continuing the (convenient and harmless) abuse of notation, we shall also use the derived binary relations given in Definition 2.2 to describe uncertainty relations between events. Thus, $E \sim F$ if and only if E and F are equally likely, and $E > F$ if and only if E is strictly more likely than F. Since, for all $c_1 < c_2$,

$$c_1 \equiv \{c_2 \,|\, \emptyset, c_1 \,|\, \Omega\} < \{c_2 \,|\, \Omega, c_1 \,|\, \emptyset\} \equiv c_2,$$

it is always true, as one would expect, that $\emptyset < \Omega$.

It is worth stressing once again at this point that *all* the order relations over $\mathcal{A} \times \mathcal{A}$, and hence over $\mathcal{C} \times \mathcal{C}$ and $\mathcal{E} \times \mathcal{E}$, are to be understood as *personal*, in the sense that, given an agreed structure for a decision problem, each individual is free to express his or her own personal preferences, in the light of his or her initial state of information M_0. Thus, for a given individual, a statement such as $E > F$ is to be interpreted as "*this individual, given the state of information described by M_0, considers event E to be more likely than event F*". Moreover, Definition 2.3 provides such a statement with an *operational meaning* since for all $c_1 < c_2$, $E > F$ is equivalent to an agreement to choose option $\{c_2 \,|\, E, c_1 \,|\, E^c\}$ in preference to option $\{c_2 \,|\, F, c_1 \,|\, F^c\}$.

To complete our discussion of basic ideas and definitions, we need to consider one further important topic. Throughout this section, we have stressed that preferences, initially defined among options but inducing binary relations among consequences and events, are *conditional* on the current state of information. The *initial* state of information, taking as an arbitrary "origin" the first occasion on which an individual thinks systematically about the problem, has been denoted by M_0. Subsequently, however, we shall need to take into account *further information*, obtained by considering the occurrence of real-world events. Given the assumed occurrence of a possible event G, preferences between options will be described by a new binary relation $\leq_G$, taking into account both the initial information M_0 and the additional information provided by G. The obvious relation between $\leq$ and $\leq_G$ is given by the following:

Definition 2.4. (***Conditional preference***). *For any* $G > \emptyset$,

(i) $a_1 \leq_G a_2 \iff$ *for all* $a\{a_1 \,|\, G, a \,|\, G^c\} \leq \{a_2 \,|\, G, a \,|\, G^c\}$;

(ii) $E \leq_G F \iff$ *for* $c_1 \leq_G c_2, \{c_2 \,|\, E, c_1 \,|\, E^c\} \leq_G \{c_2 \,|\, F, c_1 \,|\, F^c\}$.

The intuitive content of the definition is clear. If we do not prefer a_1 to a_2, given G, then this preference obviously carries over to any pair of options leading, respectively, to a_1 or a_2 if G occurs, and defined identically if G^c occurs. Conversely, comparison of options which are identical if G^c occurs depends entirely on consideration of what happens if G occurs. Naturally, the induced binary relations set out in Definition 2.2 have their obvious counterparts, denoted by $\sim_G$ and $<_G$.

The induced binary relation between consequences is obviously defined by

$$c_1 \leq_G c_2 \iff \{c_1 \,|\, \Omega\} \leq_G \{c_2 \,|\, \Omega\}.$$

However, when we come, in Section 2.3, to discuss the desirable properties of $\leq$ and $\leq_G$ we shall make formal assumptions which imply that, as one would expect, $c_1 \leq_G c_2$ if and only if $c_1 \leq c_2$, so that preferences between pure consequences are not affected by additional information regarding the uncertain events in $\mathcal{E}$.

The definition of the conditional uncertainty relation $\leq_G$ is a simple translation of Definition 2.3 to a conditional preference setting. The conditional uncertainty relation $\leq_G$ induced between events is of fundamental importance. This relation, with its derived forms $\sim_G$ and $<_G$, provides the key to investigating the way in which uncertainties about events should be modified in the light of new information. Obviously, if $G = \Omega$, all conditional relations reduce to their unconditional counterparts. Thus, it is only when $\emptyset < G < \Omega$ that conditioning on G may yield new preference patterns.

2.3 COHERENCE AND QUANTIFICATION

2.3.1 Events, Options and Preferences

The formal representation of the decision-maker's "wider frame of discourse" includes an algebra of events $\mathcal{E}$, a set of consequences $\mathcal{C}$, and a set of options $\mathcal{A}$, whose generic element has the form $\{c_j \,|\, E_j, j \in J\}$, where $\{E_j, j \in J\}$ is a finite partition of the certain event Ω, $E_j \in \mathcal{E}$, $c_j \in \mathcal{C}$, $j \in J$. The set $\mathcal{A} \times \mathcal{A}$ is equipped with a collection of binary relations $\leq_G$, $G > \emptyset$, representing the notion that one option is not preferred to another, given the assumed occurrence of a possible event G. In addition, all preferences are assumed conditional on an initial state of information, M_0, with the binary relation $\leq$ (i.e., $\leq_\Omega$) representing the preference relation on $\mathcal{A} \times \mathcal{A}$ conditional on M_0 alone.

We now wish to make precise our assumptions about these elements of the formal representation of a decision problem. Bearing in mind the overall objective of developing a rational approach to choosing among options, our assumptions, presented in the form of a series of *axioms*, can be viewed as responses to the questions: "what rules should preference relations obey?" and "what events should be included in $\mathcal{E}$?"

Each formal axiom will be accompanied by a detailed discussion of the intuitive motivation underlying it.

It is important to recognise that the axioms we shall present are *prescriptive*, not *descriptive*. Thus, they do not purport to describe the ways in which individuals actually *do* behave in formulating problems or making choices, neither do they assert, on some presumed "ethical" basis, the ways in which individuals *should* behave. The axioms simply prescribe constraints which it seems to us imperative to acknowledge in those situations where an individual aspires to choose among alternatives in such a way as to avoid certain forms of behavioural inconsistency.

2.3.2 Coherent Preferences

We shall begin by assuming that problems represented within the formal framework are non-trivial and that we are able to compare any pair of simple *dichotomised* options.

Axiom 1. (***Comparability of consequences and dichotomised options***).

(i) *There exist consequences* c_1, c_2 *such that* $c_1 < c_2$.

(ii) *For all consequences* c_1, c_2, *and events* E, F,
either $\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq \{c_2 \,|\, F, c_1 \,|\, F^c\}$
or $\{c_2 \,|\, E, c_1 \,|\, E^c\} \geq \{c_2 \,|\, F, c_1 \,|\, F^c\}$.

Discussion of Axiom 1. Condition (i) is very natural. If all consequences were equivalent, there would not be a decision problem in any real sense, since all choices would certainly lead to precisely equivalent outcomes. We have already noted that, in any given decision problem, $\mathcal{C}$ can be defined as simply the set of consequences required for that problem. Condition (ii) does *not* therefore assert that we should be able to compare *any* pair of conceivable options, however bizarre or fantastic. In most *practical* problems, there will typically be a high degree of similarity in the form of the consequences (e.g. all monetary), although it is easy to think of examples where this form is complex (e.g. combinations of monetary, health and industrial relations elements). We are trying to capture the essence of what is required for an orderly and systematic approach to comparing alternatives of genuine interest. We are not, at this stage, making the direct assumption that *all options*, however complex, can be compared. But there could be no possibility of an orderly and systematic approach if we were unwilling to express preferences among simple dichotomised options and hence (with $E = F = \Omega$) among the consequences themselves. Condition (ii) is therefore to be interpreted in the following sense: "*If we aspire to make a rational choice between alternative options, then we must at least be willing to express preferences between simple dichotomised options.*"

There are certainly many situations where we find the task of comparing simple options, and even consequences, very difficult. Resource allocation among competing health care programmes involving different target populations and morbidity and mortality rates is one obvious such example. However, the difficulty of comparing options in such cases does not, of course, obviate the *need* for such comparisons if we are to aspire to responsible decision making.

We shall now state our assumptions about the ways in which preferences should fit together or *cohere* in terms of the order relation over $\mathcal{A} \times \mathcal{A}$.

Axiom 2. (***Transitivity of preferences***).

(i) $a \leq a$.

(ii) *If* $a_1 \leq a_2$ *and* $a_2 \leq a_3$*, then* $a_1 \leq a_3$.

Discussion of Axiom 2. Condition (i) has obvious intuitive support. It would make little sense to assert that an option was strictly preferred to itself. It would also seem strangely perverse to claim to be unable to compare an option with itself! We note that, from Definition 2.2 (i), if $a \leq a$, then $a \sim a$. Condition (ii) requires preferences to be *transitive*. The intuitive basis for such a requirement is perhaps best illustrated by considering the consequences of *intransitive* preferences. Suppose, therefore, that we found ourselves expressing the preferences $a_1 < a_2$, $a_2 < a_3$ and $a_3 < a_1$ among three options a_1, a_2 and a_3. The assertion of strict preference rules out equivalence between any pair of the options, so that our expressed preferences reveal that we perceive *some actual difference in value* (no matter how small) between the two options in each case. Let us now examine the behavioural implications of these expressed preferences. If we consider, for example, the preference $a_1 < a_2$, we are implicitly stating that there exists a "price", say x, that we would be willing to pay in order to move from a position of having to accept option a_1 to one where we have, instead, to accept option a_2. Let y and z denote the corresponding "prices" for switching from a_2 to a_3 and from a_3 to a_1, respectively. Suppose now that we are confronted with the prospect of having to accept option a_1. By virtue of the expressed preference $a_1 < a_2$ and the above discussion, we are willing to pay x in order to exchange option a_1 for option a_2. But now, by virtue of the preference $a_2 < a_3$, we are willing to pay y in order to exchange a_2 for a_3. Repeating the argument once again, since $a_3 < a_1$ we are willing to pay z in order to avoid a_3 and have, instead, the prospect of option a_1. *We would thus have paid $x + y + z$ in order to find ourselves in precisely the same position as we started from!* What is more, we could find ourselves arguing through this cycle over and over again. Willingness to act on the basis of intransitive preferences is thus seen to be equivalent to a willingness to suffer unnecessarily the *certain loss* of something to which one attaches positive value. We regard this as inherently inconsistent behaviour and recall that the purpose of the axioms is to impose rules of coherence on preference orderings that will exclude the possibility of such inconsistencies. Thus, Axiom 2(ii) is to be understood in the following sense: "*If we aspire to avoid expressing preferences whose behavioural implications are such as to lead us to the certain loss of something we value, then we must ensure that our preferences fit together in a transitive manner.*"

Our discussion of this axiom is, of course, informal and appeals to directly intuitive considerations. At this stage, it would therefore be inappropriate to become involved in a formal discussion of terms such as "value" and "price". It is intuitively clear that if we assert strict preference there must be some amount of money (or grains of wheat, or beads, or whatever), however small, having a "value" less than the perceived difference in "value" between the two options. We should therefore be willing to pay this amount to switch from the less preferred to the more preferred option.

The following consequences of Axiom 2 are easily established and will prove useful in our subsequent development.

Proposition 2.1. (***Transitivity of uncertainties***).

(i) $E \sim E$.

(ii) $E_1 \leq E_2$ *and* $E_2 \leq E_3$ *imply* $E_1 \leq E_3$.

Proof. This is immediate from Definition 2.3 and Axiom 2. $\triangleleft$

Proposition 2.2. (***Derived transitive properties***).

(i) *If* $a_1 \sim a_2$ *and* $a_2 \sim a_3$ *then* $a_1 \sim a_3$.
If $E_1 \sim E_2$ *and* $E_2 \sim E_3$ *then* $E_1 \sim E_3$.

(ii) *If* $a_1 < a_2$ *and* $a_2 \sim a_3$ *then* $a_1 < a_3$.
If $E_1 < E_2$ *and* $E_2 \sim E_3$ *then* $E_1 < E_3$.

Proof. To prove (i), let $a_1 \sim a_2$ and $a_2 \sim a_3$ so that, by Definition 2.2, $a_1 \leq a_2$, $a_2 \leq a_1$ and $a_2 \leq a_3$, $a_3 \leq a_2$. Then, by Axiom 2(ii), $a_1 \leq a_3$ and $a_3 \leq a_1$, and thus $a_1 \sim a_3$. A similar argument applies to events using Proposition 2.1. Again, part (ii) follows rather similarly. $\triangleleft$

Axiom 3. (***Consistency of preferences***).

(i) *If* $c_1 \leq c_2$ *then, for all* $G > \emptyset$, $c_1 \leq_G c_2$.

(ii) *If, for some* $c_1 < c_2$, $\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq \{c_2 \,|\, F, c_1 \,|\, F^c\}$, *then* $E \leq F$.

(iii) *If, for some* c *and* $G > \emptyset$, $\{a_1 \,|\, G, c \,|\, G^c\} \leq \{a_2 \,|\, G, c \,|\, G^c\}$,
then $a_1 \leq_G a_2$.

Discussion of Axiom 3. Condition (i) formalises the idea that preferences between pure consequences should not be affected by the acquisition of further information regarding the uncertain events in $\mathcal{E}$. Conditions (ii) and (iii) ensure that Definitions 2.3 and 2.4 have operational content. Indeed, (ii) asserts that if we have $\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq \{c_2 \,|\, F, c_1 \,|\, F^c\}$ for *some* $c_1 < c_2$ then we should have this preference for *any* $c_1 < c_2$. This formalises the intuitive idea that the stated preference should only depend on the "relative likelihood" of E and F and should *not* depend on the particular consequences used in constructing the options. Similarly, (iii) asserts that if we have the preference $\{a_1 \,|\, G, c \,|\, G^c\} \leq \{a_2 \,|\, G, c \,|\, G^c\}$ for *some* c then, given G, a_1 should not be preferred to a_2, so that, for *any* a, $\{a_1 \,|\, G, a \,|\, G^c\} \leq \{a_2 \,|\, G, a \,|\, G^c\}$. This latter argument is a version of what might be called the *sure-thing principle*: if two situations are such that whatever the outcome of the first there is a preferable corresponding outcome of the second, then the second situation is preferable overall.

An important implication of Axiom 3 is that preferences between consequences are invariant under changes in the information "origin" regarding events in $\mathcal{E}$.

Proposition 2.3. (***Invariance of preferences between consequences***).
$c_1 \leq c_2$ *if and only if there exist* $G > \emptyset$ *such that* $c_1 \leq_G c_2$.

Proof. If $c_1 \leq c_2$ then, by Axiom 3(i), $c_1 \leq_G c_2$ for any event G. Conversely, by Definition 2.4(i), for any $G > \emptyset$, $c_1 \leq_G c_2$ implies that for *any* option a, one has $\{c_1 \,|\, G, a \,|\, G^c\} \leq \{c_2 \,|\, G, a \,|\, G^c\}$. Taking $a = \{c_1 \,|\, G, c_2 \,|\, G^c\}$, this implies that $\{c_1 \,|\, G, c_2 \,|\, G^c\} \leq \{c_1 \,|\, \emptyset, c_2 \,|\, \Omega\}$. If $c_1 > c_2$ this implies, by Axiom 3(ii), that $G \leq \emptyset$, thus contradicting $G > \emptyset$. Hence, by Axiom 1(ii), $c_1 \leq c_2$. $\triangleleft$

Another important consequence of Axiom 3 is that uncertainty orderings of events respect logical implications, in the sense that if E logically implies F, i.e., if $E \subseteq F$, then F cannot be considered less likely than E.

Proposition 2.4. (***Monotonicity***). *If $E \subseteq F$ then $E \leq F$.*

Proof. For any $c_1 < c_2$, define

$$a_1 = \{c_2 \,|\, E, c_1 \,|\, E^c\} = \{c_1 \,|\, F - E, \{c_2 \,|\, E, c_1 \,|\, E^c\} \,|\, (F-E)^c\},$$
$$a_2 = \{c_2 \,|\, F, c_1 \,|\, F^c\} \;= \{c_2 \,|\, F - E, \{c_2 \,|\, E, c_1 \,|\, E^c\} \,|\, (F-E)^c\}.$$

By Axiom 3(i) with $G = F - E = F \cap E^c$, $a_1 \leq a_2$. It now follows immediately from Definition 2.2 that $E \leq F$. $\triangleleft$

This last result is an example of how coherent *qualitative* comparisons of uncertain events in terms of the "not more likely" relation conform to intuitive requirements.

If follows from Proposition 2.4 that, as one would expect, for any event E, $\emptyset \leq E \leq \Omega$. We shall mostly work, however, with "significant" events, for which this ordering is strict.

Definition 2.5. (***Significant events***). *An event E is significant given $G > \emptyset$ if $c_1 <_G c_2$ implies that $c_1 <_G \{c_2 \,|\, E, c_1 \,|\, E^c\} <_G c_2$. If $G = \Omega$, we shall simply say that E is significant.*

Intuitively, significant events given G are those operationally perceived by the decision-maker as "practically possible but not certain" given the information provided by G. Thus, given $G > \emptyset$ and assuming $c_1 <_G c_2$, if E is judged to be significant given G, one would strictly prefer the option $\{c_2 \,|\, E, c_1 \,|\, E^c\}$ to c_1 for sure, since it provides an additional perceived possibility of obtaining the more desirable consequence c_2. Similarly, one would strictly prefer c_2 for sure to the stated option.

Proposition 2.5. (***Characterisation of significant events***). *An event E is significant given $G > \emptyset$, if and only if $\emptyset < E \cap G < G$. In particular, E is significant if and only if $\emptyset < E < \Omega$.*

Proof. Using Definitions 2.4 and 2.5, if E is significant given G then, for all $c_1 \leq_G c_2$ and for any option a,

$$\{c_1 \,|\, G, a \,|\, G^c\} < \{c_2 \,|\, E \cap G, c_1 \,|\, E^c \cap G, a \,|\, G^c\} < \{c_2 \,|\, G, a \,|\, G^c\}\,.$$

Taking $a = c_1$, we have

$$c_1 = \{c_2 \,|\, \emptyset, c_1 \,|\, \Omega\} < \{c_2 \,|\, E \cap G, c_1 \,|\, (E \cap G)^c\} < \{c_2 \,|\, G, c_1 \,|\, G^c\}$$

and hence, by Definition 2.3, $\emptyset < E \cap G < G$. Conversely, if $\emptyset < E \cap G < G$,

$$\{c_1 \,|\, G, c_1 \,|\, G^c\} < \{c_2 \,|\, E \cap G, c_1 \,|\, E^c \cap G, c_1 \,|\, G^c\} < \{c_2 \,|\, G, c_1 \,|\, G^c\}$$

and hence, by Axiom 3(iii), $c_1 <_G \{c_2 \,|\, E, c_1 \,|\, E^c\} <_G c_2$. If, in particular, $G = \Omega$ then E is significant if and only if $\emptyset < E < \Omega$. $\lhd$

The operational essence of "learning from experience" is that a decision-maker's preferences may change in passing from one state of information to a new state brought about by the acquisition of further information regarding the occurrence of events in $\mathcal{E}$, which leads to changes in assessments of uncertainty. There are, however, too many complex ways in which such changes in assessments can take place for us to be able to capture the idea in a simple form. On the other hand, the very special case in which preferences do *not* change is easy to describe in terms of the concepts thus far available to us.

Definition 2.6. (***Pairwise independence of events***).
We say that E and F are (pairwise) independent, denoted by $E \perp F$, if, and only if, for all c, c_1, c_2

(i) $c \bullet \{c_2 \,|\, E, c_1 \,|\, E^c\} \Rightarrow c \bullet_F \{c_2 \,|\, E, c_1 \,|\, E^c\}$,

(ii) $c \bullet \{c_2 \,|\, F, c_1 \,|\, F^c\} \Rightarrow c \bullet_E \{c_2 \,|\, F, c_1 \,|\, F^c\}$,

where $\bullet$ is any one of the relations $<$, $\sim$ or $>$.

The definition is given for the simple situation of preferences between pure consequences and dichotomised options. Since by Proposition 2.3 preferences regarding pure consequences are unaffected by additional information, the condition stated captures, in an operational form, the notion that uncertainty judgements about E, say, are unaffected by the additional information F. We interpret $E \perp F$ as "E is independent of F". An alternative characterisation will be given in Proposition 2.13.

2.3.3 Quantification

The notion of preference between options, formalised by the binary relation $\leq$, provides a *qualitative* basis for comparing options and, by extension, for comparing consequences and events. The *coherence axioms* (Axioms 1 to 3) then provide a minimal set of rules to ensure that qualitative comparisons based on $\leq$ cannot have intuitively undesirable implications.

We shall now argue that this purely qualitative framework is inadequate for serious, systematic comparisons of options. An illuminating analogy can be drawn between $\leq$ and a number of qualitative relations in common use both in an everyday setting and in the physical sciences.

Consider, for example, the relations *not heavier than, not longer than, not hotter than.* It is abundantly clear that these cannot suffice, as they stand, as an adequate basis for the physical sciences. Instead, we need to introduce in each case some form of *quantification* by setting up a *standard unit of measurement*, such as the *kilogram*, the *metre*, or the *centigrade interval*, together with an (implicitly) continuous scale such as *arbitrary decimal fractions* of a kilogram, a metre, a centigrade interval. This enables us to assign a *numerical value*, representing *weight*, *length*, or *temperature*, to any given physical or chemical entity.

This can be achieved by carrying out, implicitly or explicitly, a series of qualitative pairwise comparisons of the feature of interest with appropriately chosen points on the standard scale. For example, in quantifying the length of a stick, we place one end against the origin of a metre scale and then use a series of qualitative comparisons, based on "not longer than" (and derived relations, such as "strictly longer than"). If the stick is "not longer than" the scale mark of 2.5 metres, but is "strictly longer than" the scale mark of 2.4 metres, we might lazily report that the stick is "2.45 metres long". If we needed to, we could continue to make qualitative comparisons of this kind with finer subdivisions of the scale, thus extending the number of decimal places in our answer. The example is, of course, a trivial one, but the general point is extremely important. *Precision, through quantification, is achieved by introducing some form of numerical standard into a context already equipped with a coherent qualitative ordering relation.*

We shall regard it as essential to be able to *aspire* to some kind of quantitative precision in the context of comparing options. It is therefore necessary that we have available some form of *standard options*, whose definitions have close links with an easily understood numerical scale, and which will play a role analogous to the standard metre or standard kilogram. As a first step towards this, we make the following assumption about the algebra of events, $\mathcal{E}$.

Axiom 4. (***Existence of standard events***). *There exists a subalgebra $\mathcal{S}$ of $\mathcal{E}$ and a function $\mu : \mathcal{S} \to [0,1]$ such that:*

(i) $S_1 \leq S_2$ *if, and only if,* $\mu(S_1) \leq \mu(S_2)$*;*

(ii) $S_1 \cap S_2 = \emptyset$ *implies that* $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2)$*;*

(iii) *for any number* α *in* $[0,1]$, *and events* E, F, *there is a standard event* S *such that* $\mu(S) = \alpha$, $E \perp S$ *and* $F \perp S$*;*

(iv) $S_1 \perp S_2$ *implies that* $\mu(S_1 \cap S_2) = \mu(S_1)\mu(S_2)$.

(v) *if* $E \perp S$, $F \perp S$ *and* $E \perp F$, *then* $E \sim S \Rightarrow E \sim_F S$.

Discussion of Axiom 4. A family of events satisfying conditions (i) and (ii) is easily identified by imagining an idealised roulette wheel of unit circumference. We suppose that no point on the circumference is "favoured" as a resting place for the ball (considered as a point) in the sense that given any c_1, c_2 and events S_1, S_2 corresponding to the ball landing within specified connected arcs, or finite unions and intersections of such arcs, $\{c_1 \mid S_1, c_2 \mid S_1^c\}$ and $\{c_1 \mid S_2, c_2 \mid S_2^c\}$ are considered equivalent if and only if $\mu(S_1) = \mu(S_2)$, where μ is the function mapping the "arc-event" to its total length. Conditions (i) and (ii) are then intuitively obvious, as is the fact, in (iii), that for any $\alpha \in [0, 1]$ we can construct an S with $\mu(S) = \alpha$. Note that $\mathcal{S}$ is required to be an algebra and thus both $\emptyset$ and Ω are standard events. It follows from Proposition 2.4 and Axiom 4(i) that $\mu(\emptyset) = 0$ and $\mu(\Omega) = 1$. The remainder of (iii) is intuitively obvious; we note first that the basic idea of an idealised roulette wheel does assume that each "play" on such a wheel is "independent", in the sense of Definition 2.6, of any other events, including previous "plays" on the same wheel. Thus, for any events E, F in $\mathcal{E}$, we can always think of an "independent" play which generates independent events S in $\mathcal{S}$ with $\mu(S) = \alpha$ for any specified α in $[0, 1]$. In this extended setting, if we think of the circumferences for two independent plays as unravelled to form the sides of a unit square, with μ mapping events to the areas they define, condition (iv) is clearly satisfied. Finally, (v) encapsulates an obviously desirable consequence of independence; namely, that if E is independent of F and S, and F is independent of S, a judgement of equivalence between E and S should not be affected by the occurrence of F.

We will refer to $\mathcal{S}$ as a *standard family* of events in $\mathcal{E}$ and will think of $\mathcal{E}$ as the algebra generated by the relevant events in the decision problem together with the elements of $\mathcal{S}$. Other forms of standard family satisfying (i) to (v) are easily imagined. For example, it is obvious that a roulette wheel of unit circumference could be imagined cut at some point and "unravelled" to form a unit interval. The underlying image would then be that of a point landing in the unit interval and an event S such that $\mu(S) = p$ would denote a subinterval of length p; alternatively, we could imagine a point landing in the unit square, with S denoting a region of area p. The obvious intuitive content of conditions (i) to (v) can clearly be similarly motivated in these cases, the discussion for the unit interval being virtually identical to that given for the roulette wheel. It is important to emphasise that we do *not* require the assumption that standard families of events actually, physically exist, or could be *precisely* constructed in accordance with conditions (i) to (v). We only require that we can invoke such a set up as a mental image.

There is, of course, an element of mathematical idealisation involved in thinking about *all* $p \in [0, 1]$, rather than, for example, some subset of the rationals, corresponding to binary expansions consisting of zeros from some specified location onwards, reflecting the inherent limits of accuracy in any actual procedure for determining arc lengths or areas. The same is true, however, of *all* scientific discourse in which measurements are taken, in principle, to be real numbers, rather than a subset of the rationals chosen to reflect the limits of accuracy in the physical measurement procedure being employed. Our argument for accepting this degree of mathematical idealisation in setting up our formal system is the same as would apply in the physical sciences. Namely, that no serious conceptual distortion is introduced, while many irrelevant technical difficulties are avoided; in particular, those concerning the non-closure of a set of numbers with respect to operations of interest. This argument is not universally accepted, however, and further, related discussion of the issue is provided in Section 2.8.

Our view is that, from the perspective of the foundations of decision-making, the step from the finite to the infinite implicit in making use of real numbers is simply a pragmatic convenience, whereas the step from comparing a finite set of possibilities to comparing an infinite set has more substantive implications. We have emphasised this latter point by postponing infinite extensions of the decision framework until Chapter 3.

Proposition 2.6. (***Collections of disjoint standard events***).
For any finite collection $\{\alpha_1, \ldots, \alpha_n\}$ *of real numbers such that* $\alpha_i > 0$ *and* $\alpha_1 + \cdots + \alpha_n \leq 1$ *there exists a corresponding collection* $\{S_1, \ldots, S_n\}$ *of disjoint standard events such that* $\mu(S_i) = \alpha_i$, $i = 1, \ldots, n$.

Proof. By Axiom 4(iii) there exists S_1 such that $\mu(S_1) = \alpha_1$. For $1 < j \leq n$, suppose inductively that $S_1, \ldots, S_{j-1}$ are disjoint, $B_j = S_1 \cup \cdots \cup S_{j-1}$ and define $\beta_j = \alpha_1 + \cdots + \alpha_{j-1} = \mu(B_j)$. By Axiom 4 (iii, iv), there exists T_j in $\mathcal{S}$ such that $\mu(B_j \cap T_j) = \mu(B_j)\{\alpha_j/(1-\beta_j)\}$. Define $S_j = T_j \cap B_j^c$, so that $S_j \cap S_i = \emptyset$, $i = 1, \ldots, j-1$. Then, $T_j = S_j \cup (T_j \cap B_j)$ and hence, using Axiom 4(ii), $\mu(T_j) = \mu(S_j) + \mu(T_j \cap B_j)$. Thus, $\mu(S_j) = \alpha_j/(1-\beta_j) - \alpha_j\beta_j/(1-\beta_j) = \alpha_j$ and the result follows. $\triangleleft$

Axiom 5. (***Precise measurement of preferences and uncertainties***).

(i) *If* $c_1 \leq c \leq c_2$, *there exists a standard event* S *such that* $c \sim \{c_2 \mid S, c_1 \mid S^c\}$.

(ii) *For each event* E, *there exists a standard event* S *such that* $E \sim S$.

Discussion of Axiom 5. In the introduction to this section, we discussed the idea of precision through quantification and pointed out, using analogies with other measurement systems such as weight, length and temperature, that the process is based on successive comparisons with a standard. Let S_q denote a standard event such that $\mu(S_q) = q$. We start with the obvious preferences, $\{c_2 \mid S_0, c_1 \mid S_0^c\} \leq c \leq \{c_2 \mid S_1, c_1 \mid S_1^c\}$, for any $c_1 \leq c \leq c_2$, and then begin to explore comparisons with standard options based on S_x, S_y with $0 < x < y < 1$. In this way, by gradually increasing x away from 0 and decreasing y away from 1, we arrive at comparisons such as $\{c_2 \mid S_x, c_1 \mid S_x^c\} \leq c \leq \{c_2 \mid S_y, c_1 \mid S_y^c\}$, with the difference $y - x$ becoming increasingly small. Intuitively, as we increase x, $\{c_2 \mid S_x, c_1 \mid S_x^c\}$ becomes more and more "attractive" as an option, and as we decrease y, $\{c_2 \mid S_y, c_1 \mid S_y^c\}$ becomes less "attractive". Any given consequence c, such that $c_1 \leq c \leq c_2$, can therefore be "sandwiched" arbitrarily tightly and, in the limit, be judged equivalent to one of the standard options defined in terms of c_1, c_2. The essence of Axiom 5(i) is that we can proceed to a common limit, α, say, approached from below by the successive values of x and above by the successive values of y. The standard family of options is thus assumed to provide a continuous scale against which any consequence can be *precisely* compared.

Condition (ii) extends the idea of precise comparison to include the assumption that, for any event E and for all consequences c_1, c_2 such that $c_1 < c_2$, the option $\{c_2 \mid E, c_1 \mid E^c\}$ can be compared precisely with the family of standard options $\{c_2 \mid S_x, c_1 \mid S_x^c\}, x \in [0, 1]$, defined by c_1 and c_2. The underlying idea is similar to that motivating condition (i). Indeed, given the intuitive content of the relation "not more likely than", we can begin with the obvious ordering $\{c_2 \mid S_0, c_1 \mid S_0^c\} \leq \{c_2 \mid E, c_1 \mid E^c\} \leq \{c_2 \mid S_1, c_1 \mid S_1^c\}$ for any event E, and then consider refinements of this of the form $\{c_2 \mid S_x, c_1 \mid S_x^c\} \leq \{c_2 \mid E, c_1 \mid E^c\} \leq \{c_2 \mid S_y, c_1 \mid S_y^c\}$, with x increasing gradually from 0, y decreasing gradually from 1, and $y - x$ becoming increasingly small, so that, in terms of the ordering of the events, $S_x \leq E \leq S_y$. Again, the essence of the axiom is that this "sandwiching" can be refined arbitrarily closely by an increasing sequence of x's and a decreasing sequence of y's tending to a common limit.

The preceding argument certainly again involves an element of mathematical idealisation. In practice, there might, in fact, be some *interval of indifference*, in the sense that we judge $\{c_2 \mid S_x, c_1 \mid S_x^c\} \leq c \leq \{c_2 \mid S_y, c_1 \mid S_y^c\}$ for some (possibly rational) x and y but feel unable to express a more precise form of preference. This is analogous to the situation where a physical measuring instrument has inherent limits, enabling one to conclude that a reading is in the range 3.126 to 3.135, say, but not permitting a more precise statement. In this case, we would typically report the measurement to be 3.13 and proceed *as if* this were a precise measurement. We formulate the theory on the prescriptive assumption that we aspire to exact measurement (exact comparisons in our case), whilst acknowledging that, in practice, we have to make do with the best level of precision currently available (or devote some resources to improving our measuring instruments!).

In the context of measuring beliefs, several authors have suggested that this imprecision be *formally incorporated into the axiom system.* For many applications, this would seem to be an unnecessary confusion of the *prescriptive* and the *descriptive*. Every physicist or chemist knows that there are inherent limits of accuracy in any given laboratory context but, so far as we know, no one has suggested developing the structures of theoretical physics or chemistry on the assumption that quantities appearing in fundamental equations should be constrained to take values in some subset of the rationals. However, it may well be that there are situations where imprecision in the context of comparing consequences is too basic and problematic a feature to be adequately dealt with by an approach based on theoretical precision, tempered with pragmatically acknowledged approximation. We shall return to this issue in Section 2.8.

The particular standard option to which c is judged equivalent will, of course, depend on c, but we have implicitly assumed that it does *not* depend on any information we might have concerning the occurrence of real-world events. Indeed, Proposition 2.3 implies that our "attitudes" or "values" regarding consequences are *fixed* throughout the analysis of any particular decision problem. It is intuitively obvious that, if the time-scale on which values change were not rather long compared with the time-scale within which individual problems are analysed, there would be little hope for rational analysis of any kind.

2.4 BELIEFS AND PROBABILITIES

2.4.1 Representation of Beliefs

It is clear that an individual's preferences among options in any decision problem should depend, at least in part, on the "degrees of belief" which that individual attaches to the uncertain events forming part of the definitions of the options.

The principles of *coherence* and *quantification* by comparison with a standard, expressed in axiomatic form in the previous section, will enable us to give a formal definition of *degree of belief*, thus providing a numerical measure of the uncertainty attached to each event.

The conceptual basis for this numerical measure will be seen to derive from the formal rules governing quantitative, coherent preferences, irrespective of the nature of the uncertain events under consideration. This is in vivid contrast to what are sometimes called the *classical* and *frequency* approaches to defining numerical measures of uncertainty (see Section 2.8), where the existence of *symmetries* and the possibility of *indefinite replication*, respectively, play fundamental roles in defining the concepts for restricted classes of events.

We cannot emphasise strongly enough the important distinction between *defining a general concept* and *evaluating a particular case*. Our *definition* will depend only on the logical notions of quantitative, coherent preferences; our practical *evaluations* will often make use of perceived symmetries and observed frequencies.

We begin by establishing some basic results concerning the uncertainty relation between events.

Proposition 2.7. (***Complete comparability of events***).
Either $E_1 > E_2$, or $E_1 \sim E_2$, or $E_2 > E_1$.

Proof. By Axiom 5(ii), there exist S_1 and S_2 such that $E_1 \sim S_1$ and $E_2 \sim S_2$; the complete ordering now follows from Axiom 4(i) and Proposition 2.1. $\triangleleft$

We see from Proposition 2.7 that, although the order relation $\leq$ between options was not assumed to be complete (i.e., not *all* pairs of options were assumed to be comparable), it turns out, as a consequence of Axiom 5 (the axiom of precise measurement), that the uncertainty relation induced between events *is* complete. A similar result concerning the comparability of all options will be established in Section 2.5.

Proposition 2.8. (***Additivity of uncertainty relations***). *If $A \leq B, C \leq D$ and $A \cap C = B \cap D = \emptyset$, then $A \cup C \leq B \cup D$. Moreover, if $A < B$ or $C < D$, then $A \cup C < B \cup D$.*

Proof. We first show that, for any G, if $A \cap G = B \cap G = \emptyset$ then $A \leq B \iff A \cup G \leq B \cup G$. For any $c_2 > c_1$, $A \cap G = B \cap G = \emptyset$, define:

$$\begin{aligned}
a_1 &= \{c_2 \,|\, A, c_1 \,|\, A^c\} = \{c_1 \,|\, G, \{c_2 \,|\, A, c_1 \,|\, A^c\} \,|\, G^c\} \\
a_2 &= \{c_2 \,|\, B, c_1 \,|\, B^c\} = \{c_1 \,|\, G, \{c_2 \,|\, B, c_1 \,|\, B^c\} \,|\, G^c\} \\
a_3 &= \{c_2 \,|\, A \cup G, c_1 \,|\, (A \cup G)^c\} = \{c_2 \,|\, G, \{c_2 \,|\, A, c_1 \,|\, A^c\} \,|\, G^c\} \\
a_4 &= \{c_2 \,|\, B \cup G, c_1 \,|\, (B \cup G)^c\} = \{c_2 \,|\, G, \{c_2 \,|\, B, c_1 \,|\, B^c\} \,|\, G^c\}.
\end{aligned}$$

Then, by Definition 2.3, $A \leq B \iff a_1 \leq a_2$; by Axiom 3, $a_1 \leq a_2 \iff a_3 \leq a_4$; and using again Definition 2.3, $a_3 \leq a_4 \iff A \cup G \leq B \cup G$. Thus,

$$A \cup (C - B) \leq B \cup (C - B) = B \cup C = C \cup (B - C) \leq D \cup (B - C),$$

$$A \cup C = A \cup (C - B) \cup (C \cap B) \leq D \cup (B - C) \cup (C \cap B) = B \cup D.$$

The final statement follows from essentially the same argument. $\triangleleft$

We now make the key definition which enables us to move to a *quantitative* notion of degree of belief.

Definition 2.7. (***Measure of degree of belief***). *Given an uncertainty relation $\leq$, the* ***probability*** *$P(E)$ of an event E is the real number $\mu(S)$ associated with any standard event S such that $E \sim S$.*

This definition provides a natural, operational extension of the qualitative uncertainty relation encapsulated in Definition 2.3, by linking the equivalence of any $E \in \mathcal{E}$ to some $S \in \mathcal{S}$ and exploiting the fact that the nature of the construction of $\mathcal{S}$ provides a direct obvious quantification of the uncertainty regarding S.

With our operational definition, the *meaning* of a probability statement is clear. For instance, the statement $P(E) = 0.5$ precisely means that E is judged to be equally likely as a standard event of 'measure' 0.5, maybe a conceptual perfect coin falling heads, or a computer generated 'random' integer being an odd number.

It should be emphasised that, according to Definition 2.7, probabilities are always *personal degrees of belief*, in that they are a numerical representation of the decision-maker's personal uncertainty relation $\leq$ between events. Moreover, probabilities are always conditional on the information currently available. It makes no sense, within the framework we are discussing, to qualify the word probability with adjectives such as "objective", "correct" or "unconditional".

Since probabilities are obviously conditional on the initial state of information M_0, a more precise and revealing notation in Definition 2.7 would have been $P(E \mid M_0)$. In order to avoid cumbersome notation, we shall stick to the shorter version, but the implicit conditioning on M_0 should always be borne in mind.

Proposition 2.9. (***Existence and uniqueness***). *Given an uncertainty relation $\leq$, there exists a unique probability $P(E)$ associated with each event E.*

Proof. Existence follows from Axiom 5(ii). For uniqueness, if $E \sim S_1$ and $E \sim S_2$ then by Proposition 2.2(ii), $S_1 \sim S_2$. The result now follows from Axiom 4(i). $\triangleleft$

Definition 2.8. (***Compatibility***). *A function $f : \mathcal{E} \to \Re$ is said to be compatible with an order relation $\leq$ on $\mathcal{E} \times \mathcal{E}$ if, for all events,*

$$E \leq F \iff f(E) \leq f(F).$$

Proposition 2.10. (***Compatibility of probability and degrees of belief***). *The probability function $P(.)$ is compatible with the uncertainty relation $\leq$.*

Proof. By Axiom 5(ii) there exist standard events S_1 and S_2 such that $E \sim S_1$ and $F \sim S_2$. Then, by Proposition 2.2(ii) , $E \leq F$ iff $S_1 \leq S_2$ and hence, by Axiom 4(i), iff $\mu(S_1) \leq \mu(S_2)$. The result follows from Definition 2.7. $\triangleleft$

The following proposition is of fundamental importance. It establishes that coherent, quantitative degrees of belief have the structure of a finitely additive probability measure over $\mathcal{E}$. Moreover, it establishes that significant events, i.e., events which are "practically possible but not certain", should be assigned probability values in the *open* interval $(0, 1)$.

Proposition 2.11. (***Probability structure of degrees of belief***).

(i) $P(\emptyset) = 0$ *and* $P(\Omega) = 1$.

(ii) *If* $E \cap F = \emptyset$*, then* $P(E \cup F) = P(E) + P(F)$.

(iii) E *is significant if, and only if,* $0 < P(E) < 1$.

Proof. (i) By Definition 2.7, $0 \leq P(E) \leq 1$. Moreover, by Axiom 4(iii) there exist S_* and S^* such that $\mu(S_*) = 0$ and $\mu(S^*) = 1$. By Proposition 2.4, $\emptyset \leq S_*$ and, by Proposition 2.10 $P(\emptyset) \leq 0$; hence, $P(\emptyset) = 0$; similarly, $S^* \leq \Omega$ implies that $P(\Omega) = 1$.

(ii) If $E = \emptyset$ or $F = \emptyset$, or both, the result is trivially true. If $E > \emptyset$ and $F > \emptyset$, then, by Proposition 2.8, $E \cup F > E$; thus, if $\alpha = P(E)$ and $\beta = P(E \cup F)$, we have $\alpha < \beta$ and, by Proposition 2.6, there exist events S_1, S_2 such that $S_1 \cap S_2 = \emptyset$, $P(S_1) = \alpha$ and $P(S_2) = \beta - \alpha$. By Proposition 2.7, $F > S_2$ or $F \sim S_2$ or $F < S_2$. If $F > S_2$, then, by Proposition 2.8, $E \cup F > S_1 \cup S_2$ and hence $P(E \cup F) > \beta$, which is impossible; similarly, if $F < S_2$ then $E \cup F < S_1 \cup S_2$ and $P(E \cup F) < \beta$ which, again, is impossible. Hence, $F \sim S_2$ and therefore $P(F) = \beta - \alpha$, so that $P(E \cup F) = P(E) + P(F)$, as stated.

(iii) By Proposition 2.5, E is significant iff $\emptyset < E < \Omega$. The result then follows immediately from Proposition 2.10. ◁

Corollary. (***Finitely additive structure of degrees of belief***).

(i) *If $\{E_j, j \in J\}$ is a finite collection of disjoint events, then*

$$P\Big(\bigcup_{j \in J} E_j\Big) = \sum_{j \in J} P(E_j).$$

(ii) *For any event E, $P(E^c) = 1 - P(E)$.*

Proof. The first part follows by induction from Proposition 2.11(iii); the second part is a special case of (i) since if $\cup_j E_j = \Omega$ then, by Proposition 2.11(i), $\Sigma_j P(E_j) = 1$. ◁

Proposition 2.11 is crucial. It establishes formally that coherent, quantitative measures of uncertainty about events must take the form of probabilities, therefore justifying the nomenclature adopted in Definition 2.6 for this measure of degree of belief. In short, *coherent degrees of belief are probabilities*.

It will often be convenient for us to use probability terminology, without explicit reference to the fact that the mathematical structure is merely serving as a representation of (personal) degrees of belief. The latter fact should, however, be constantly borne in mind.

Definition 2.9. (***Probability distribution***). *If $\{E_j, j \in J\}$ form a finite partition of Ω, with $P(E_j) = p_j, j \in J$, then $\{p_j, j \in J\}$ is said to be a probability distribution over the partition.*

This terminology will prove useful in later discussions. The idea is that total belief (in Ω, having measure 1) is *distributed* among the events of the partition, $\{E_j, j \in J\}$, according to the relative degrees of belief $\{p_j, j \in J\}$, with $\Sigma_j p_j = \Sigma_j P(E_j) = 1$.

Starting from the qualitative ordering among events, we have derived a quantitative measure, $P(.) \equiv P(.\,|\,M_0)$, over $\mathcal{E}$ and shown that, expressed in conventional mathematical terminology, it has the form of a *finitely additive probability measure*, compatible with the qualitative ordering $\leq$. We now establish that this is the only probability measure over $\mathcal{E}$ compatible with $\leq$.

Proposition 2.12. (***Uniqueness of the probability measure***). *P is the only probability measure compatible with the uncertainty relation $\leq$.*

Proof. If P' were another compatible measure, then by Proposition 2.8 we would always have $P'(E) \leq P'(F) \iff P(E) \leq P(F)$; hence, there exists a monotonic function f of $[0, 1]$ into itself such that $P'(E) = f\{P(E)\}$. By Proposition 2.6, for all non-negative α, β such that $\alpha + \beta \leq 1$, there exist disjoint standard events S_1 and S_2, such that $P(S_1) = \alpha$ and $P(S_2) = \beta$. Hence, by Axiom 4(ii), $f(\alpha + \beta) = P'(S_1 \cup S_2) = P'(S_1) + P'(S_2) = f(\alpha) + f(\beta)$ and so (Eichhorn, 1978, Theorem 2.63), $f(\alpha) = k\alpha$ for all α in $[0, 1]$. But, by Proposition 2.9, $P'(\Omega) = 1$ and hence, $k = 1$, so that we have $P'(E) = P(E)$ for all E. $\triangleleft$

We shall now establish that our operational definition of (pairwise) independence of events is compatible with its more standard, *ad hoc*, product definition.

Proposition 2.13. (***Characterisation of independence***).

$$E \perp F \iff P(E \cap F) = P(E)P(F).$$

Proof. Suppose $E \perp F$. By Axiom 4(iii), there exists S_1 such that $P(S_1) = P(E)$, $E \perp S_1$ and $F \perp S_1$. Hence, by Axiom 4(v), $E \sim_F S_1$, so that, for any consequences $c_1 < c_2$, and any option a,

$$\{c_2\,|\,E \cap F, c_1\,|\,E^c \cap F, a\,|\,F^c\} \sim \{c_2\,|\,S_1 \cap F, c_1\,|\,S_1^c \cap F, a\,|\,F^c\}.$$

Taking $a = c_1$, we have

$$\{c_2\,|\,E \cap F, c_1\,|\,(E \cap F)^c\} \sim \{c_2\,|\,S_1 \cap F, c_1\,|\,(S_1 \cap F)^c\},$$

so that $E \cap F \sim S_1 \cap F$. Again by Axiom 4(iii), given F, S_1, there exists S_2 such that $P(S_2) = P(F)$, $F \perp S_2$ and $S_1 \perp S_2$. Hence, by an identical argument to the above, and noting from Definition 2.6 the symmetry of $\perp$, we have

$$S_1 \cap F \sim S_1 \cap S_2.$$

By Propositions 2.1, 2.10, and Axiom 4(iv),

$$P(E \cap F) = P(S_1 \cap S_2) = P(S_1)P(S_2),$$

and hence $P(E \cap F) = P(E)P(F)$.

Suppose $P(E \cap F) = P(E)P(F)$. By Axiom 4(iii), there exists S such that $P(S) = P(F)$ and $F \perp S$, $E \perp S$. Hence, by the first part of the proof,

$$P(E \cap S) = P(E)P(S) = P(E)P(F) = P(E \cap F),$$

so that $E \cap F \sim E \cap S$. Now suppose, without loss of generality, that $c \leq \{c_2 \,|\, E, c_1 \,|\, E^c\}$. Then, by Definition 2.6,

$$\{c \,|\, S, c_1 \,|\, S^c\} \leq \{c_2 \,|\, E \cap S, c_1 \,|\, (E \cap S)^c\}.$$

But $\{c \,|\, S, c_1 \,|\, S^c\} \sim \{c \,|\, F, c_1 \,|\, F^c\}$ and

$$\{c_2 \,|\, E \cap S, c_1 \,|\, (E \cap S)^c\} \sim \{c_2 \,|\, E \cap F, c_1 \,|\, (E \cap F)^c\};$$

hence by Proposition 2.2,

$$\{c \,|\, F, c_1 \,|\, F^c\} \leq \{c_2 \,|\, E \cap F, c_1 \,|\, (E \cap F)^c\},$$

so that $c \leq_F \{c_2 \,|\, E, c_1 \,|\, E^c\}$. A similar argument can obviously be given reversing the roles of E and F, hence establishing that $E \perp F$. $\triangleleft$

2.4.2 Revision of Beliefs and Bayes' Theorem

The assumed occurrence of a real-world event will typically modify preferences between options by modifying the degrees of belief attached, by an individual, to the events defining the options. In this section, we use the assumptions of Section 2.3 in order to identify the precise way in which coherent modification of initial beliefs should proceed.

The starting point for analysing order relations between events, given the assumed occurrence of a possible event G, is the uncertainty relation $\leq_G$ defined between events. Given the assumed occurrence of $G > \emptyset$, the ordering $\leq$ between acts is replaced by $\leq_G$. Analogues of Propositions 2.1 and 2.2 are trivially established and we recall (Proposition 2.3) that, for any $G > \emptyset$, $c_2 \leq c_1$ iff $c_2 \leq_G c_1$.

Proposition 2.14. (***Properties of conditional beliefs***).

(i) $E \leq_G F \iff E \cap G \leq F \cap G$.

(ii) *If there exist* $c_1 < c_2$ *such that* $\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq_G \{c_2 \,|\, F, c_1 \,|\, F^c\}$, *then* $E \leq_G F$.

Proof. By Definition 2.4 and Proposition 2.3, $E \leq_G F$ iff, for all $c_2 \geq c_1$,

$$\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq_G \{c_2 \,|\, F, c_1 \,|\, F^c\},$$

i.e., if, and only if, for all a,

$$\{c_2 \,|\, E \cap G, c_1 \,|\, E^c \cap G, a \,|\, G^c\} \leq \{c_2 \,|\, F \cap G, c_1 \,|\, F^c \cap G, a \,|\, G^c\}.$$

Taking $a = c_1$,

$$E \leq_G F \iff \{c_2 \,|\, E \cap G, c_1 \,|\, (E \cap G)^c\} \leq \{c_2 \,|\, F \cap G, c_1 \,|\, (F \cap G)^c\},$$

and this is true iff $E \cap G \leq F \cap G$.

Moreover, if there exist $c_2 > c_1$ such that $\{c_2 \,|\, E, c_1 \,|\, E^c\} \leq_G \{c_2 \,|\, F, c_1 \,|\, F^c\}$ then, by Definition 2.4, with $a = c_1$,

$$\{c_2 \,|\, E \cap G, c_1 \,|\, E^c \cap G, c_1 \,|\, G^c\} \leq \{c_2 \,|\, F \cap G, c_1 \,|\, F^c \cap G, c_1 \,|\, G^c\},$$

so that

$$\{c_2 \,|\, E \cap G, c_1 \,|\, (E \cap G)^c\} \leq \{c_2 \,|\, F \cap G, c_1 \,|\, (F \cap G)^c\}$$

and the result follows from Axiom 3(ii) and part (i) of this proposition. $\triangleleft$

Definition 2.10. (***Conditional measure of degree of belief***). *Given a conditional uncertainty relation $\leq_G, G > \emptyset$, the conditional probability $P(E \,|\, G)$ of an event E given the assumed occurrence of G is the real number $\mu(S)$ such that $E \sim_G S$, where S is an standard event independent of G.*

Generalising the idea encapsulated in Definition 2.7, $P(E \,|\, G)$ provides a quantitative operational measure of the uncertainty attached to E given the assumed occurrence of the event G. The following fundamental result provides the key to the process of revising beliefs in a coherent manner in the light of new information. It relates the *conditional* measure of degree of belief $P(. \,|\, G)$ to the *initial* measure of degree of belief $P(.)$.

We have, of course, already stressed that *all* degrees of belief are conditional. The intention of the terminology used above is to emphasise the additional conditioning resulting from the occurrence of G; the initial state of information, M_0, is always present as a conditioning factor, although omitted throughout for notational convenience.

Proposition 2.15. (***Conditional probability***). *For any $G > \emptyset$,*

$$P(E \,|\, G) = \frac{P(E \cap G)}{P(G)} \cdot$$

Proof. By Axiom 4(iii) and Proposition 2.13, there exists $S \perp G$ such that $\mu(S) = P(E \cap G)/P(G)$. By Proposition 2.13,

$$P(S \cap G) = P(S)P(G) = \mu(S)P(G) = P(E \cap G).$$

Thus, by Proposition 2.10, $S \cap G \sim E \cap G$ and, by Proposition 2.14, $S \sim_G E$. Thus, by Definition 2.10, $P(E \,|\, G) = \mu(S) = P(E \cap G)/P(G)$. $\triangleleft$

Note that, in our formulation, $P(E\,|\,G) = P(E \cap G)/P(G)$ is a logical derivation from the axioms, *not* an *ad hoc* definition. In fact, this is the simplest version of *Bayes' theorem*. An extended form is given later in Proposition 2.19.

Proposition 2.16. (***Compatibility of conditional probability and conditional degrees of belief***).

$$E \leq_G F \iff P(E\,|\,G) \leq P(F\,|\,G).$$

Proof. By Proposition 2.14(i), $E \leq_G F$ iff $E \cap G \leq F \cap G$, which, by Proposition 2.10, holds if and only if $P(E \cap G) \leq P(F \cap G)$; the result now follows from Proposition 2.15. $\triangleleft$

We now extend Proposition 2.11 to degrees of belief conditional on the occurrence of significant events.

Proposition 2.17. (***Probability structure of conditional degrees of belief***). *For any event $G > \emptyset$,*

(i) $P(\emptyset\,|\,G) = 0 \leq P(E\,|\,G) \leq P(\Omega\,|\,G) = 1$*;*

(ii) *if $E \cap F \cap G = \emptyset$, then $P(E \cup F\,|\,G) = P(E\,|\,G) + P(F\,|\,G)$;*

(iii) *E is significant given G* $\iff 0 < P(E\,|\,G) < 1$.

Proof. By Proposition 2.15, $P(E\,|\,G) \geq 0$ and $P(\emptyset\,|\,G) = 0$; moreover, since $E \cap G \leq G$, Proposition 2.10 implies that $P(E \cap G) \leq P(G)$, so that, by Proposition 2.15, $P(E\,|\,G) \leq 1$. Finally, $\Omega \cap G = G$, so that, using again Proposition 2.15 , $P(\Omega\,|\,G) = 1$.

By Proposition 2.15,

$$\begin{aligned} P(E \cup F\,|\,G) &= \frac{P\big((E \cap G) \cup (F \cap G)\big)}{P(G)} \\ &= \frac{P(E \cap G)}{P(G)} + \frac{P(F \cap G)}{P(G)} = P(E\,|\,G) + P(F\,|\,G). \end{aligned}$$

Finally, by Proposition 2.5, E is significant given G iff $\emptyset < E \cap G < G$. Thus, by Proposition 2.10, E is significant given G iff $0 < P(E \cap G) < P(G)$. The result follows from Proposition 2.15. $\triangleleft$

Corollary. (***Finitely additive structure of conditional degrees of belief***). *For all $G > \emptyset$,*

(i) *if $\{E_j \cap G, j \in J\}$ is a finite collection of disjoint events, then*

$$P\Big(\bigcup_{j \in J} E_j \,\Big|\, G\Big) = \sum_{j \in J} P(E_j\,|\,G);$$

(ii) *for any event E, $P(E^c\,|\,G) = 1 - P(E\,|\,G)$.*

Proof. This parallels the proof of the Corollary to Proposition 2.11. $\triangleleft$

Proposition 2.18. (***Uniqueness of the conditional probability measure***). *$P(.\,|\,G)$ is the only probability measure compatible with the conditional uncertainty relation $\leq_G$.*

Proof. This parallels the proof of Proposition 2.12. $\triangleleft$

Example 2.1. ***(Simpson's paradox)***. The following example provides an instructive illustration of the way in which the formalism of conditional probabilities provides a coherent resolution of an otherwise seemingly paradoxical situation.

Suppose that the results of a clinical trial involving 800 sick patients are as shown in Table 2.1, where T, T^c denote, respectively, that patients did or did not receive a certain treatment, and R, R^c denote, respectively, that the patients did or did not recover.

Table 2.1 *Trial results for all patients*

	R	R^c	*Total*	*Recovery rate*
T	200	200	400	50%
T^c	160	240	400	40%

Intuitively, it seems clear that the treatment is beneficial, and were one to base probability judgements on these reported figures, it would seem reasonable to specify

$$P(R\,|\,T) = 0.5, \qquad P(R\,|\,T^c) = 0.4,$$

where recovery and the receipt of treatment by individuals are now represented, in an obvious notation, as events. Suppose now, however, that one became aware of the trial outcomes for male and female patients separately, and that these have the summary forms described in Tables 2.2 and 2.3.

Table 2.2 *Trial results for male patients*

	R	R^c	*Total*	*Recovery rate*
T	180	120	300	60%
T^c	70	30	100	70%

The results surely seem paradoxical. Tables 2.2 and 2.3 tell us that the treatment is neither beneficial for males nor for females; but Table 2.1 tells us that overall it is beneficial! How are we to come to a coherent view in the light of this apparently conflicting evidence?

Table 2.3 *Trial results for female patients*

	R	R^c	*Total*	*Recovery rate*
T	20	80	100	20%
T^c	90	210	300	30%

The seeming paradox is easily resolved by an appeal to the logic of probability which, after all, we have just demonstrated to be the prerequisite for the coherent treatment of uncertainty. With M, M^c denoting, respectively, the events that a patient is either male or female, were one to base probability judgements on the figures reported in Tables 2.2 and 2.3, it would seem reasonable to specify

$$P(R\,|\,M\cap T) = 0.6, \qquad P(R\,|\,M\cap T^c) = 0.7\,,$$

$$P(R\,|\,M^c\cap T) = 0.2, \qquad P(R\,|\,M^c\cap T^c) = 0.3\,.$$

To see that these judgements do indeed cohere with those based on Table 2.1, we note, from the Corollary to Proposition 2.11, Proposition 2.15 and the Corollary to Proposition 2.17, that

$$P(R\,|\,T) = P(R\,|\,M\cap T)P(M\,|\,T) + P(R\,|\,M^c\cap T)P(M^c\,|\,T)$$
$$P(R\,|\,T^c) = P(R\,|\,M\cap T^c)P(M\,|\,T^c) + P(R\,|\,M^c\cap T^c)P(M^c\,|\,T^c),$$

where

$$P(M\,|\,T) = 0.75, \qquad P(M\,|\,T^c) = 0.25.$$

The probability formalism reveals that the seeming paradox has arisen from the confounding of sex with treatment as a consequence of the unbalanced trial design. See Simpson (1951), Blyth (1972, 1973) and Lindley and Novick (1981) for further discussion.

Proposition 2.19. (***Bayes' theorem***).
For any finite partition $\{E_j, j\in J\}$ of Ω and $G > \emptyset$,

$$P(E_i\,|\,G) = \frac{P(G\,|\,E_i)P(E_i)}{\sum_{j\in J} P(G\,|\,E_j)P(E_j)}\,.$$

Proof. By Proposition 2.15,

$$P(E_i\,|\,G) = \frac{P(E_i\cap G)}{P(G)} = \frac{P(G\,|\,E_i)P(E_i)}{P(G)}\,.$$

The result now follows from the Corollary to Proposition 2.11 when applied to $G = \cup_j(G\cap E_j)$. $\triangleleft$

Bayes' theorem is a simple mathematical consequence of the fact that quantitative coherence implies that degrees of belief should obey the rules of probability. From another point of view, it may also be established (Zellner, 1988b) that, under some reasonable desiderata, Bayes' theorem is an optimal information processing system.

Since the $\{E_j, j \in J\}$ form a partition and hence, by the Corollary to Proposition 2.17, $\sum_j P(E_j \,|\, G) = 1$, Bayes' theorem may be written in the form

$$P(E_j \,|\, G) \propto P(G \,|\, E_j)P(E_j), \quad j \in J,$$

since the missing proportionality constant is $[P(G)]^{-1} = [\Sigma_j P(G \,|\, E_j)P(E_j)]^{-1}$, and thus it is always possible to normalise the products by dividing by their sum. This form of the theorem is often very useful in applications.

Bayes' theorem acquires a particular significance in the case where the uncertain events $\{E_j, j \in J\}$ correspond to an exclusive and exhaustive set of *hypotheses* about some aspect of the world (for example, in a medical context, the set of possible diseases from which a patient may be suffering) and the event G corresponds to a relevant piece of *evidence*, or *data* (for example, the outcome of a clinical test). If we adopt the more suggestive notation, $E_j = H_j, j \in J, G = D$, and, as usual, we omit explicit notational reference to the initial state of information M_0, Proposition 2.17 leads to Bayes' theorem in the form $P(H_j \,|\, D) = P(D \,|\, H_j)P(H_j)/P(D), j \in J$, where $P(D) = \Sigma_j P(D \,|\, H_j)P(H_j)$, characterizing the way in which initial beliefs about the hypotheses, $P(H_j)$, $j \in J$, are modified by the data, D, into a revised set of beliefs, $P(H_j \,|\, D), j \in J$. This process is seen to depend crucially on the specification of the quantities $P(D \,|\, H_j)$, $j \in J$, which reflect how beliefs about obtaining the given data, D, vary over the different underlying hypotheses, thus defining the "relative likelihoods" of the latter. The four elements, $P(H_j)$, $P(D \,|\, H_j)$, $P(H_j \,|\, D)$ and $P(D)$, occur, in various guises, throughout Bayesian statistics and it is convenient to have a standard terminology available.

Definition 2.11. (***Prior, posterior, and predictive probabilities***).
If $\{H_j, j \in J\}$ are exclusive and exhaustive events (hypotheses), then for any event (data) D,

(i) *$P(H_j)$, $j \in J$, are called the* ***prior probabilities*** *of the H_j, $j \in J$;*

(ii) *$P(D \,|\, H_j)$, $j \in J$, are called the* ***likelihoods*** *of the H_j, $j \in J$, given D;*

(iii) *$P(H_j \,|\, D)$, $j \in J$, are called the* ***posterior probabilities*** *of the H_j, $j \in J$;*

(iv) *$P(D)$ is called the* ***predictive probability*** *of D implied by the likelihoods and the prior probabilities.*

It is important to realise that the terms "prior" and "posterior" only have significance *given* an initial state of information and *relative* to an additional piece of information. Thus, $P(H_j)$, which could be more properly be written as $P(H_j \,|\, M_0)$, represents beliefs prior to conditioning on data D, but posterior to conditioning on whatever history led to the state of information described by M_0. Similarly, $P(H_j \mid D)$, or, more properly, $P(H_j \mid M_0 \cap D)$, represents beliefs posterior to conditioning on M_0 and D, but prior to conditioning on any further data which may be obtained subsequent to D.

The predictive probability $P(D)$, logically implied by the likelihoods and the prior probabilities, provides a basis for assessing the compatibility of the data D with our beliefs (see Box, 1980). We shall consider this in more detail in Chapter 6.

Example 2.2. ***(Medical diagnosis)***. In simple problems of medical diagnosis, Bayes' theorem often provides a particularly illuminating form of analysis of the various uncertainties involved. For simplicity, let us consider the situation where a patient may be characterised as belonging either to state H_1, or to state H_2, representing the presence or absence, respectively, of a specified disease. Let us further suppose that $P(H_1)$ represents *the prevalence rate* of the disease in the population to which the patient is assumed to belong, and that further information is available in the form of the result of a single clinical test, whose outcome is either positive (suggesting the presence of the disease and denoted by $D = T$), or negative (suggesting the absence of the disease and denoted by $D = T^c$).

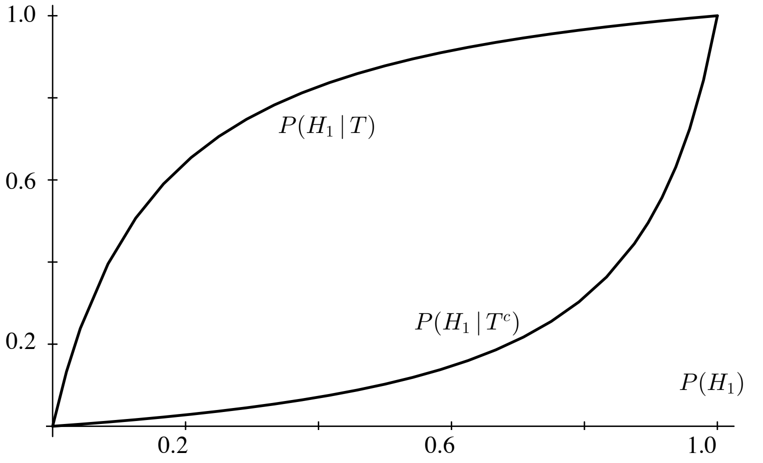

Figure 2.2 *$P(H_1 \mid T)$ and $P(H_1 \mid T^c)$ as functions of $P(H_1)$*

The quantities $P(T \mid H_1)$ and $P(T^c \mid H_2)$ represent the *true positive* and *true negative rates* of the clinical test (often referred to as the *test sensitivity* and *test specificity*, respectively) and the systematic use of Bayes' theorem then enables us to understand the manner in which these characteristics of the test combine with the prevalence rate to produce varying degrees of diagnostic discriminatory power. In particular, for a given clinical test of known sensitivity and specificity, we can investigate the range of underlying prevalence rates for which the test has worthwhile diagnostic value.

As an illustration of this process, let us consider the assessment of the diagnostic value of stress thallium-201 scintigraphy, a technique involving analysis of Gamma camera image data as an indicator of coronary heart disease. On the basis of a controlled experimental study, Murray *et al.* (1981) concluded that $P(T \mid H_1) = 0.900$, $P(T^c \mid H_2) = 0.875$ were reasonable orders of magnitude for the sensitivity and specificity of the test.

Insight into the diagnostic value of the test can be obtained by plotting values of $P(H_1 \mid T)$, $P(H_1 \mid T^c)$ against $P(H_1)$, where

$$P(H_1 \mid D) = \frac{P(D \mid H_1)P(H_1)}{P(D \mid H_1)P(H_1) + P(D \mid H_2)P(H_2)} \ ,$$

for $D = T$ or $D = T^c$, as shown in Figure 2.2.

As a single, overall measure of the discriminatory power of the test, one may consider the difference $P(H_1 \mid T) - P(H_1 \mid T^c)$. In cases where $P(H_1)$ has very low or very high values (e.g. for large population screening or following individual patient referral on the basis of suspected coronary disease, respectively), there is limited diagnostic value in the test. However, in clinical situations where there is considerable uncertainty about the presence of coronary heart disease, for example, $0.25 \leq P(H_1) \leq 0.75$, the test may be expected to provide valuable diagnostic information.

One further point about the terms prior and posterior is worth emphasising. *They are not necessarily to be interpreted in a chronological sense*, with the assumption that "prior" beliefs are specified first and then later modified into "posterior" beliefs. Propositions 2.15 and 2.17 do not involve any such chronological notions. They merely indicate that, *for coherence*, specifications of degrees of belief must satisfy the given relationships. Thus, for example, in Proposition 2.15 one might first specify $P(G)$ and $P(E \mid G)$ and then use the relationship stated in the theorem to arrive at coherent specification of $P(E \cap G)$. In any given situation, the particular order in which we specify degrees of belief and check their coherence is a pragmatic one; thus, some assessments seem straightforward and we feel comfortable in making them directly, while we are less sure about other assessments and need to approach them indirectly via the relationships implied by coherence. It is true that the natural order of assessment does coincide with the "chronological" order in a number of practical applications, but it is important to realise that this is a pragmatic issue and not a requirement of the theory.

2.4.3 Conditional Independence

An important special case of Proposition 2.15 arises when E and G are such that $P(E \mid G) = P(E)$, so that beliefs about E are *unchanged* by the assumed occurrence of G. Not surprisingly, this is directly related to our earlier operational definition of (pairwise) independence.

Proposition 2.20. *For all* $F > \emptyset$, $E \perp F \iff P(E \mid F) = P(E)$.

Proof. $E \perp F \iff P(E \cap F) = P(E)P(F)$ and, by Proposition 2.15, we have $P(E \cap F) = P(E \mid F)P(F)$. $\triangleleft$

In the case of three events, E, F and G, the situation is somewhat more complicated in that, from an intuitive point of view, we would regard our degree of belief for E as being "independent" of knowledge of F and G if and only if $P(E \mid H) = P(E)$, for any of the four possible forms of H,

$$\{F \cap G,\ F^c \cap G,\ F \cap G^c,\ F^c \cap G^c\},$$

describing the combined occurrences, or otherwise, of F and G (and, of course, similar conditions must hold for the "independence" of F from E and G, and of G from E and F). These considerations motivate the following formal definition, which generalises Definition 2.6 and can be shown (see e.g. Feller, 1950/1968, pp. 125–128) to be necessary and sufficient for encapsulating, in the general case, the intuitive conditions discussed above.

Definition 2.12. (***Mutual independence***).
Events $\{E_j, j \in J\}$ are said to be mutually independent if, for any $I \subseteq J$,

$$P\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} P(E_i).$$

An important consequence of the fact that coherent degrees of belief combine in conformity with the rules of (finitely additive) mathematical probability theory is that the task of specifying degrees of belief for complex combinations of events is often greatly simplified. Instead of being forced into a direct specification, we can attempt to represent the complex event in terms of simpler events, for which we feel more comfortable in specifying degrees of belief. The latter are then recombined, using the probability rules, to obtain the desired specification for the complex event. Definition 2.12 makes clear that the judgement of independence for a collection of events leads to considerable additional simplification when complex intersections of events are to be considered. Note that Proposition 2.20 derives from the uncertainty relation $\leq_F$ and therefore reflects an inherently personal judgement (although coherence may rule out some events from being judged independent: for example, any E, F such that $\emptyset \subset E \subseteq F \subset \Omega$).

There is a sense, however, in which the judgement of independence (given M_0) for large classes of events of interest reflects a rather extreme form of belief, in that scope for learning from experience is very much reduced. This motivates consideration of the following weaker form of independence judgement.

Definition 2.13. (***Conditional independence***). *The events $\{E_j, j \in J\}$ are said to be conditionally independent given $G > \emptyset$ if, for any $I \subseteq J$,*

$$P\left(\bigcap_{i \in I} E_i \,|\, G\right) = \prod_{i \in I} P(E_i \,|\, G).$$

For any subalgebra $\mathcal{F}$ of $\mathcal{E}$, the events $\{E_j, j \in J\}$ are said to be conditionally independent given $\mathcal{F}$ if and only if they are conditionally independent given any $G > \emptyset$ in $\mathcal{F}$.

Definitions 2.12 and 2.13 could, of course, have been stated in primitive terms of choices among options, as in Definition 2.6. However, having seen in detail the way in which the latter leads to the standard "product definition", it will be clear that a similar equivalence holds in these more general cases, but that the algebraic manipulations involved are somewhat more tedious.

The form of degree of belief judgement encapsulated in Definition 2.13 is one which is utilised in some way or another in a wide variety of practical contexts and statements of scientific theories. Indeed, a detailed discussion of the kinds of circumstances in which it may be reasonable to structure beliefs on the basis of such judgements will be a main topic of Chapter 4. Thus, for example, in the practical context of sampling, with or without replacement, from large dichotomised populations (of voters, manufactured items, or whatever), successive outcomes (voting intention, marketable quality, . . .) may very often be judged independent, *given exact knowledge of the proportional split in the dichotomised population.* Similarly, in simple Mendelian theory, the genotypes of successive offspring are typically judged to be independent events, *given the knowledge of the two genotypes forming the mating.* In the absence of such knowledge, however, in neither case would the judgement of independence for successive outcomes be intuitively plausible, since earlier outcomes provide information about the unknown population or mating composition and this, in turn, influences judgements about subsequent outcomes. For a detailed analysis of the concept of conditional independence, see Dawid (1979a, 1979b, 1980b).

2.4.4 Sequential Revision of Beliefs

Bayes' theorem characterises the way in which current beliefs about a set of mutually exclusive and exhaustive hypotheses, H_j, $j \in J$, are revised in the light of new data, D. In practice, of course, we typically receive data in successive stages, so that the process of revising beliefs is sequential.

As a simple illustration of this process, let us suppose that data are obtained in two stages, which can be described by real-world events D_1 and D_2. Omitting, for convenience, explicit conditioning on M_0, revision of beliefs on the basis of the first piece of data D_1 is described by $P(H_j \,|\, D_1) = P(D_1 \,|\, H_j)P(H_j)/P(D_1)$, $j \in J$. When it comes to the further, subsequent revision of beliefs in the light of D_2, the likelihoods and prior probabilities to be used in Bayes' theorem are now $P(D_2 \,|\, H_j \cap D_1)$ and $P(H_j \,|\, D_1)$, $j \in J$, respectively, since all judgements are now conditional on D_1. We thus have, for all $j \in J$,

$$P(H_j \,|\, D_1 \cap D_2) = \frac{P(D_2 \,|\, H_j \cap D_1)P(H_j \,|\, D_1)}{P(D_2 \,|\, D_1)} \; ,$$

where $P(D_2 \,|\, D_1) = \sum_j P(D_2 \,|\, H_j \cap D_1)P(H_j \,|\, D_1)$.

From an intuitive standpoint, we would obviously anticipate that coherent revision of initial belief in the light of the combined data, $D_1 \cap D_2$, should not depend on whether D_1, D_2 were analysed successively or in combination. This is easily verified by substituting the expression for $P(H_j \,|\, D_1)$ into the expression for $P(H_j \,|\, D_1 \cap D_2)$, whereupon we obtain

$$\frac{P(D_2 \,|\, H_j \cap D_1)P(D_1 \,|\, H_j)P(H_j)}{P(D_2 \,|\, D_1)P(D_1)} = \frac{P(D_1 \cap D_2 \,|\, H_j)P(H_j)}{P(D_1 \cap D_2)} \; ,$$

the latter being the direct expression for $P(H_j \,|\, D_1 \cap D_2)$ from Bayes' theorem when $D_1 \cap D_2$ is treated as a single piece of data.

The generalisation of this sequential revision process to any number of stages, corresponding to data, $D_1, D_2, \ldots, D_n, \ldots$, proceeds straightforwardly. If we write $D^{(k)} = D_1 \cap D_2 \cap \cdots \cap D_k$ to denote all the data received up to and including stage k, then, for all $j \in J$,

$$P(H_j \,|\, D^{(k+1)}) = \frac{P(D_{k+1} \,|\, H_j \cap D^{(k)})P(H_j \,|\, D^{(k)})}{P(D_{k+1} \,|\, D^{(k)})} \; ,$$

which provides a recursive algorithm for the revision of beliefs.

There is, however, a potential practical difficulty in implementing this process, since there is an implicit need to specify the successively *conditioned likelihoods*, $P(D_{k+1} \,|\, H_j \cap D^{(k)})$, $j \in J$, a task which, in the absence of simplifying assumptions, may appear to be impossibly complex if k is at all large. One possible form of simplifying assumption is the judgement of conditional independence for $D_1, D_2, \ldots, D_n$, given any $H_j, j \in J$, since, by Definition 2.13, we then only need the evaluations $P(D_{k+1} \,|\, H_j \cap D^{(k)}) = P(D_{k+1} \,|\, H_j)$, $j \in J$. Another possibility might be to assume a rather weak form of dependence by making the judgement that a (Markov) property such as $P(D_{k+1} \,|\, H_j \cap D^{(k)}) = P(D_{k+1} \,|\, H_j \cap D_k)$, $j \in J$, holds for all k. As we shall see later, these kinds of simplifying structural assumptions play a fundamental role in statistical modelling and analysis.

In the case of two hypotheses, H_1, H_2, the judgement of conditional independence for $D_1, D_2, \ldots, D_n, \ldots$, given H_1 or H_2, enables us to provide an alternative description of the process of revising beliefs by noting that, in this case,

$$\frac{P(H_1 \mid D^{(k+1)})}{P(H_2 \mid D^{(k+1)})} = \frac{P(H_1 \mid D^{(k)})}{P(H_2 \mid D^{(k)})} \times \frac{P(D_{k+1} \mid H_1)}{P(D_{k+1} \mid H_2)}\,.$$

With due regard to the relative nature of the terms prior and posterior, we can thus summarise the learning process (in "favour" of H_1) as follows:

$$\textit{posterior odds} = \textit{prior odds} \times \textit{likelihood ratio}.$$

In Section 2.6, we shall examine in more detail the key role played by the sequential revision of beliefs in the context of complex, sequential decision problems.

2.5 ACTIONS AND UTILITIES

2.5.1 Bounded Sets of Consequences

At the beginning of Section 2.4, we argued that choices among options are governed, in part, by the relative degrees of belief that an individual attaches to the uncertain events involved in the options. It is equally clear that choices among options should depend on the relative values that an individual attaches to the consequences flowing from the events. The measurement framework of Axiom 5(i) provides us with a direct, intuitive way of introducing a *numerical measure of value for consequences*, in such a way that the latter has a coherent, operational basis. Before we do this, we need to consider a little more closely the nature of the set of consequences $\mathcal{C}$. The following special case provides a useful starting point for our development of a measure of value for consequences.

Definition 2.14. (***Extreme consequences***). *The pair of consequences c_* and c^* are called, respectively, the* ***worst*** *and the* ***best*** *consequences in a decision problem if, for any other consequence $c \in \mathcal{C}$, $c_* \leq c \leq c^*$.*

It could be argued that *all* real decision problems actually have extreme consequences. Indeed, we recall that all consequences are to be thought of as relevant consequences in the context of the decision problem. This eliminates pathological, mathematically motivated choices of $\mathcal{C}$, which could be constructed in such a way as to rule out the existence of extreme consequences. For example, in *mathematical* modelling of decision problems involving monetary consequences, $\mathcal{C}$ is often taken to be the real line $\Re$ or, in a no-loss situation with current assets k, to be the interval $[k, \infty)$. Such $\mathcal{C}$'s would not contain *both* a best and a worst consequence but, on the other hand, they clearly do not correspond to concrete, practical problems. In the next section, we shall consider the solution to decision problems for which extreme consequences are assumed to exist.

Nevertheless, despite the force of the pragmatic argument that extreme consequences always exist, it must be admitted that insisting upon problem formulations which satisfy the assumption of the existence of extreme consequences can sometimes lead to rather tedious complications of a conceptual or mathematical nature.

Consider, for example, a medical decision problem for which the consequences take the form of different numbers of years of remaining life for a patient. Assuming that more value is attached to longer survival, it would appear rather difficult to justify any *particular* choice of realistic upper bound, even though we believe there to be one. To choose a particular c^* would be tantamount to putting forward c^* years as a realistic possible survival time, but regarding $c^* + 1$ years as impossible! In such cases, it is attractive to have available the possibility, for conceptual and mathematical convenience, of dealing with sets of consequences *not* possessing extreme elements (and the same is true of many problems involving monetary consequences). For this reason, we shall also deal (in Section 2.5.3) with the situation in which extreme consequences are not assumed to exist.

2.5.2 Bounded Decision Problems

Let us consider a decision problem $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq)$ for which extreme consequences $c_* < c^*$ are assumed to exist. We shall refer to such decision problems as *bounded*.

Definition 2.15. (***Canonical utility function for consequences***). *Given a preference relation $\leq$, the utility $u(c) = u(c \,|\, c_*, c^*)$ of a consequence c, relative to the extreme consequences $c_* < c^*$, is the real number $\mu(S)$ associated with any standard event S such that $c \sim \{c^* \,|\, S, c_* \,|\, S^c\}$. The mapping $u : \mathcal{C} \to \Re$ is called the **utility function**.*

It is important to note that the definition of utility only involves comparison among consequences and options constructed with standard events. Since the preference patterns among consequences is unaffected by additional information, we would expect the utility of a consequence to be uniquely defined and to remain unchanged as new information is obtained. This is indeed the case.

Proposition 2.21. (***Existence and uniqueness of bounded utilities***). *For any bounded decision problem $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq)$ with extreme consequences $c_* < c^*$,*

(i) *for all c, $u(c \,|\, c_*, c^*)$ exists and is unique;*

(ii) *the value of $u(c \,|\, c_*, c^*)$ is unaffected by the assumed occurrence of an event $G > \emptyset$;*

(iii) $0 = u(c_* \,|\, c_*, c^*) \leq u(c \,|\, c_*, c^*) \leq u(c^* \,|\, c_*, c^*) = 1.$

Proof. (i) Existence follows immediately from Axiom 5(i). For uniqueness, note that if $c \sim \{c^* \,|\, S_1, c_* \,|\, S_1^c\}$ and $c \sim \{c^* \,|\, S_2, c_* \,|\, S_2^c\}$ then, by transitivity and Axiom 3(ii), $\{c^* \,|\, S_1, c_* \,|\, S_1^c\} \sim \{c^* \,|\, S_2, c_* \,|\, S_2^c\}$ and $S_1 \sim S_2$; the result now follows from Axiom 4(i).

(ii) To establish this, let $c \sim \{c^* \,|\, S_1, c_* \,|\, S_1^c\}$, so that $u(c \,|\, c_*, c^*) = \mu(S_1)$; using Axiom 4(iii), for any $G > \emptyset$ choose S_2 such that $G \perp S_2$ and $\mu(S_2) = \mu(S_1)$. Then, by Definition 2.6, $c \sim_G \{c^* \,|\, S_2, c_* \,|\, S_2^c\}$ and so the utility of c given G is just the original value $\mu(S_2)$.

(iii) Finally, since $c^* \equiv \{c^* \,|\, \emptyset, c_* \,|\, \Omega\}, c_* \equiv \{c^* \,|\, \Omega, c_* \,|\, \emptyset\}$, and both $\emptyset$ and Ω belong to the algebra of standard events, we have $u(c_* \,|\, c_*, c^*) = \mu(\emptyset) = 0$ and $u(c^* \,|\, c_*, c^*) = \mu(\Omega) = 1$. It then follows, from Definition 2.15 and Axiom 4(i), that $0 \leq u(c \,|\, c_*, c^*) \leq 1$. $\triangleleft$

It is interesting to note that $u(c \,|\, c_*, c^*)$, which we shall often simply denote by $u(c)$, can be given an operational interpretation in terms of degrees of belief. Indeed, if we consider a choice between the fixed consequence c and the option $\{c^* \,|\, E, c_* \,|\, E^c\}$, for some event E, then the utility of c can be thought of as defining a threshold value for the degree of belief in E, in the sense that values greater than u would lead an individual to prefer the uncertain option, whereas values less than u would lead the individual to prefer c for certain. The value u itself corresponds to indifference between the two options and is the degree of belief in the occurrence of the best, rather than worst, consequence.

This suggests one possible technique for the experimental *elicitation of utilities*, a subject which has generated a large literature (with contributions from economists and psychologists, as well as from statisticians). We shall illustrate the ideas in Example 2.3.

Using the coherence and quantification principles set out in Section 2.3, we have seen how numerical measures can be assigned to two of the elements of a decision problem in the form of *degrees of belief for events* and *utilities for consequences*. It remains now to investigate how an *overall numerical measure of value* can be attached to an *option*, whose form depends both on the events of a finite partition of the certain event Ω and on the particular consequences to which these events lead.

Definition 2.16. (***Conditional expected utility***).
For any $c_* < c^*, G > \emptyset$, and $a \equiv \{c_j \,|\, E_j, j \in J\}$,

$$\overline{u}(a \,|\, c_*, c^*, G) = \sum_{j \in J} u(c_j \,|\, c_*, c^*) P(E_j \,|\, G)$$

is the expected utility of the option a, given G, with respect to the extreme consequences c_*, c^*. If $G = \Omega$, we shall simply write $\overline{u}(a \,|\, c_*, c^*)$ in place of $\overline{u}(a \,|\, c_*, c^*, \Omega)$.

In the language of mathematical probability theory (see Chapter 3), if the utility value of a is considered as a "random quantity", contingent on the occurrence of a particular event E_j, then $\overline{u}$ is simply the *expected value* of that utility when the probabilities of the events are considered conditional on G.

Proposition 2.22. (***Decision criterion for a bounded decision problem***). *For any bounded decision with extreme consequences $c_* < c^*$, and $G > \emptyset$,*

$$a_1 \leq_G a_2 \iff \overline{u}(a_1 \,|\, c_*, c^*, G) \leq \overline{u}(a_2 \,|\, c_*, c^*, G).$$

Proof. Let $a_i = \{c_{ij} \,|\, E_{ij}, j = 1, \ldots, n_i\}$, $i = 1, 2$. By Axioms 5(ii), 4(iii), and Proposition 2.13, for all (i, j) there exist S_{ij} and S'_{ij} such that

$$c_{ij} \sim \{c^* \,|\, S'_{ij}, c_* \,|\, S'^{c}_{ij}\}, \quad S_{ij} \perp (E_{ij} \cap G), \quad P(S'_{ij}) = P(S_{ij}).$$

Hence, by Proposition 2.10, $c_{ij} \sim \{c^* \,|\, S_{ij}, c_* \,|\, S^c_{ij}\}$ with $S_{ij} \perp (E_{ij} \cap G)$ and $P(S_{ij}) = u(c_{ij} \,|\, c_*, c^*)$. By Definition 2.6, for $i = 1, 2$ and any option a,

$$\begin{aligned}&\{[c_{ij} \,|\, E_{ij} \cap G], j = 1, \ldots, n_i,\ a \,|\, G^c\}\\ &\qquad\qquad \sim \{[(c^* \,|\, S_{ij}, c_* \,|\, S^c_{ij}) \,|\, E_{ij} \cap G], j = 1, \ldots, n_i,\ a \,|\, G^c\},\end{aligned}$$

which may be written as $\{c^* \,|\, A_i, c_* \,|\, B_i, a \,|\, G^c\}$, where $A_i = \cup_j (E_{ij} \cap G \cap S_{ij})$ and $B_i = \cup_j (E_{ij} \cap G \cap S^c_{ij})$. By Propositions 2.14(ii) and 2.16, and using Definition 2.5, $a_1 \leq_G a_2 \Rightarrow A_1 \leq_G A_2 \Rightarrow P(A_1 \,|\, G) \leq P(A_2 \,|\, G)$. But, by Proposition 2.15, $P(E_{ij} \cap G \cap S_{ij}) = P(E_{ij} \cap G) P(S_{ij}) = P(S_{ij}) P(E_{ij} \,|\, G) P(G)$. Hence,

$$P(A_i \,|\, G) = \sum_{j=1}^{n_i} u(c_{ij} \,|\, c_*, c^*) P(E_{ij} \,|\, G) = \overline{u}(a_i \,|\, c_*, c^*, G)$$

and so $a_1 \leq_G a_2 \Leftrightarrow \overline{u}(a_1 \,|\, c_*, c^*, G) \leq \overline{u}(a_2 \,|\, c, c^*, G)$. ◁

The result just established is sometimes referred to as the *principle of maximising expected utility*. In our development, this is clearly not an independent "principle", but rather an implication of our assumptions and definitions. In summary form, the resulting prescription for quantitative, coherent decision-making is: *choose the option with the greatest expected utility*.

Technically, of course, Proposition 2.22 merely establishes, for each $\leq_G$, a complete ordering of the options considered and does not guarantee the *existence* of an optimal option for which the expected utility is a maximum. However, in most (if not all) concrete, practical problems the set of options considered will be finite and so a *best option* (not necessarily unique) will exist. In more abstract mathematical formulations, the existence of a maximum will depend on analytic features of the set of options and on the utility function $u : \mathcal{C} \to \Re$.

Example 2.3. ***(Utilities of oil wildcatters)***. One of the earliest reported systematic attempts at the quantification of utilities in a practical decision-making context was that of Grayson (1960), whose decision-makers were oil wildcatters engaged in exploratory searches for oil and gas. The consequences of drilling decisions and their outcomes are ultimately changes in the wildcatters' monetary assets, and Grayson's work focuses on the assessment of utility functions for this latter quantity.

For the purposes of illustration, suppose that we restrict attention to changes in monetary assets ranging, in units of one thousand dollars, from -150 (the *worst consequence*) to $+825$ (the *best consequence*). Assuming $u(-150) = 0$, $u(825) = 1$, the above development suggests ways in which we might try to elicit an individual wildcatter's values of $u(c)$ for various c in the range $-150 < c < 825$. For example, one could ask the wildcatter, using a series of values of c, which option he or she would prefer out of the following:

(i) c for sure,

(ii) entry into a venture having outcome 825 with probability p and an outcome -150 with probability $1 - p$, for some specified p.

If c_p emerges from such interrogation as an approximate "indifference" value, the theory developed above suggests that, for a coherent individual,

$$u(c_p) = p\,u(825) + (1 - p)\,u(-150) = p.$$

Repeating this exercise for a range of values of p, provides a series of (c_p, p) pairs, from which a "picture" of $u(c)$ over the range of interest can be obtained. An alternative procedure, of course, would be to fix c, perform an interrogation for various p until an "indifference" value, p_c is found, and then repeat this procedure for a range of values of c to obtain a series of (c, p_c) pairs.

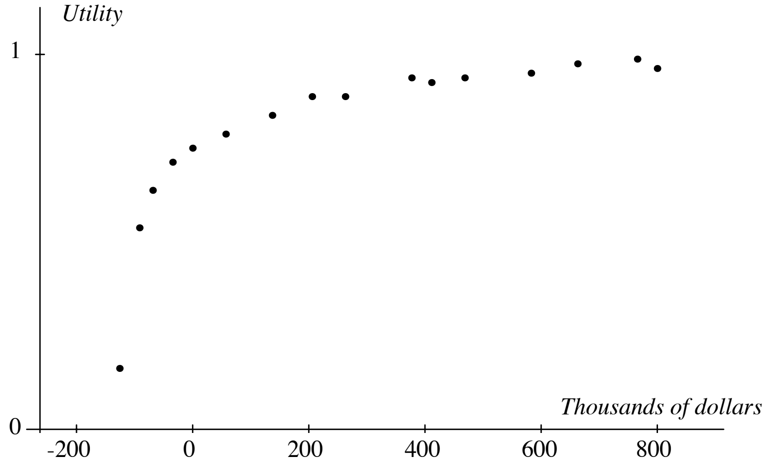

Figure 2.3 *William Beard's utility function for changes in monetary assets*

Figure 2.3 shows the results obtained by Grayson using procedures of this kind to interrogate oil company executive, W. Beard, on October 23, 1957. A "picture" of Beard's utility function clearly emerges from the empirical data. In particular, over the range concerned, the utility function reflects considerable risk aversion, in the sense that even quite small asset losses lead to large (negative) changes in utility compared with the (positive) changes associated with asset gains.

Since the expected utility $\overline{u}$ is a linear combination of values of the utility function, Proposition 2.22 guarantees that preferences among options are invariant under changes in the origin and scale of the utility measure used; i.e., invariant with respect to transformations of the form $Au(.) + B$, provided we take $A > 0$, so that the orientation of "best" and "worst" is not changed. In general, therefore, such an origin and scale can be chosen for convenience in any given problem, and we can simply refer to the expected utility of an option without needing to specify the (positive linear) transformation of the utility function which has been used. However, there may be bounded decision problems where the probabilistic interpretation discussed above makes it desirable to work in terms of *canonical* utilities, derived by referring to the best and worst consequences.

In the next section, we shall provide an extension of these ideas to more general decision problems where extreme consequences are not assumed to exist.

2.5.3 General Decision Problems

We begin with a more general definition of the utility of a consequence which preserves the linear combination structure and the invariance discussed above.

Definition 2.17. (***General utility function***). *Given a preference relation $\leq$, the utility $u(c \,|\, c_1, c_2)$ of a consequence c, relative to the consequences $c_1 < c_2$, is defined to be the real number u such that*

if $c < c_1$ and $c_1 \sim \{c_2 \,|\, S_x, c \,|\, S_x^c\}$, then $u = -x/(1-x)$;

if $c_1 \leq c \leq c_2$ and $c \sim \{c_2 \,|\, S_x, c_1 \,|\, S_x^c\}$, then $u = x$;

if $c > c_2$ and $c_2 \sim \{c \,|\, S_x, c_1 \,|\, S_x^c\}$, then $u = 1/x$

where $x = \mu(S_x)$ is the measure associated with the standard event S_x.

Our restricted definition of utility (Definition 2.15) relied on the existence of extreme consequences c_*, c^*, such that $c_* \leq c \leq c^*$ for all $c \in \mathcal{C}$. In the absence of this assumption, we have to select some *reference consequences*, c_1, c_2 to play the role of c_*, c^*. However, we cannot then assume that $c_1 \leq c \leq c_2$ for all c, and this means that if c_1, c_2 are to define a utility scale by being assigned values $0, 1$, respectively, we shall require *negative* assignments for $c < c_1$ and assignments *greater than one* for $c > c_2$. The definition is motivated by a desire to maintain the linear features of the utility function obtained in the case where extreme consequences exist. It can be checked straightforwardly that if $c_{(1)}$, $c_{(2)}$, $c_{(3)}$ denote any permutation of c, c_1, c_2, where $c_1 < c_2$ and $c_{(1)} \leq c_{(2)} \leq c_{(3)}$, the definition given ensures that for any $G > \emptyset$, $c_{(2)} \sim_G \{c_{(3)} \,|\, S_x, c_{(1)} \,|\, S_x^c\}$ implies that

$$u(c_{(2)} \,|\, c_1, c_2) = x\, u(c_{(3)} \,|\, c_1, c_2) + (1-x)\, u(c_{(1)} \,|\, c_1, c_2).$$

The following result extends Proposition 2.21 to the general utility function defined above.

Proposition 2.23. (***Existence and uniqueness of utilities***). *For any decision problem, and for any pair of consequences $c_1 < c_2$,*

(i) *for all c, $u(c \,|\, c_1, c_2)$ exists and is unique;*

(ii) *the value of $u(c \,|\, c_1, c_2)$ is unaffected by the occurrence of an event $G > \emptyset$;*

(iii) $u(c_1 \,|\, c_1, c_2) = 0$ *and* $u(c_2 \,|\, c_1, c_2) = 1$.

Proof. This is virtually identical to the proof of Proposition 2.21. $\triangleleft$

The following results guarantee that the utilities of consequences are linearly transformed if the pair of consequences chosen as a reference is changed.

Proposition 2.24. (***Linearity***). *For all $c_1 < c_2$ and $c_3 < c_4$ there exist $A > 0$ and B such that, for all c, $u(c \,|\, c_1, c_2) = Au(c \,|\, c_3, c_4) + B$.*

Proof. Suppose first that $c_3 \geq c_1$, $c_4 \leq c_2$, and $c_1 \leq c \leq c_2$. By Axiom 5(ii), $c_3 \leq c \leq c_4$ implies that there exists a standard event S_x such that $c \sim \{c_4 \,|\, S_x, c_3 \,|\, S_x^c\}$. Hence, by Proposition 2.22,

$$u(c \,|\, c_1, c_2) = xu(c_4 \,|\, c_1, c_2) + (1-x)u(c_3 \,|\, c_1, c_2),$$

where $x = P(S_x)$ and, by Definition 2.17, $u(c \,|\, c_3, c_4) = x$. Hence, $u(c \,|\, c_1, c_2) = Au(c \,|\, c_3, c_4) + B$, where $A = u(c_4 \,|\, c_1, c_2) - u(c_3 \,|\, c_1, c_2)$ and $B = u(c_3 \,|\, c_1, c_2)$.

By Axiom 5(ii), if $c_3 > c$ there exists S_y such that $c_3 \sim \{c_4 \,|\, S_y, c \,|\, S_y^c\}$. Hence, by Proposition 2.22,

$$u(c_3 \,|\, c_1, c_2) = yu(c_4 \,|\, c_1, c_2) + (1-y)u(c \,|\, c_1, c_2),$$

where $y = P(S_y)$ and, by Definition 2.17, $u(c \,|\, c_3, c_4) = -y/(1-y)$. Hence, $u(c \,|\, c_1, c_2) = Au(c \,|\, c_3, c_4) + B$, with A and B as above. Similarly, if $c > c_4$ there exists S_z such that $c_4 \sim \{c \,|\, S_z, c_3 \,|\, S_z^c\}$ and

$$u(c_4 \,|\, c_1, c_2) = yu(c \,|\, c_1, c_2) + (1-y)u(c_3 \,|\, c_1, c_2),$$

where $y = P(S_y)$ and, by Definition 2.17, $u(c \,|\, c_3, c_4) = 1/y$. Hence, we have $u(c \,|\, c_1, c_2) = Au(c \,|\, c_3, c_4) + B$, with A and B as above.

Now suppose that the c's have arbitrary order, subject to $c_2 > c_1$, $c_4 > c_3$. Let c_*, c^* be the minimum and maximum, respectively, of $\{c_1, c_2, c_3, c_4, c\}$. Then, by the above, there exist A_1, B_1, A_2, B_2 such that, for $c_{(i)} \in \{c_1, c_2, c_3, c_4, c\}$, $u(c_{(i)} \,|\, c_*, c^*) = A_1 u(c_{(i)} \,|\, c_1, c_2) + B_1$ and $u(c_{(i)} \,|\, c_*, c^*) = A_2 u(c_{(i)} \,|\, c_3, c_4) + B_2$; hence, $u(c_{(i)} \,|\, c_1, c_2) = (A_2/A_1)u(c_{(i)} \,|\, c_3, c_4) + (B_2 - B_1)/A_1$. $\triangleleft$

Finally, we generalise Proposition 2.22 to unbounded decision problems;

Proposition 2.25. (***General decision criterion***).
For any decision problem, pair of consequences $c_1 < c_2$, and event $G > \emptyset$,

$$a_1 \leq_G a_2 \iff \overline{u}(a_1 \,|\, c_1, c_2, G) \leq \overline{u}(a_2 \,|\, c_1, c_2, G).$$

Proof. Suppose $a_i = \{c_{ij} \,|\, E_{ij}, j = 1, \ldots, n_i\}, i = 1, 2$, and let c_*, c^* be such that for all c_{ij}, $c_* \leq c_{ij} \leq c^*$. Then, by Proposition 2.22, $a_2 \leq_G a_1$ iff $\overline{u}(a_2 \,|\, c_*, c^*, G) \leq \overline{u}(a_1 \,|\, c_*, c^*, G)$. But, by Proposition 2.24, there exists $A > 0$ and B such that $u(c \,|\, c_*, c^*) = Au(c \,|\, c_1, c_2) + B$, and so the result follows. $\triangleleft$

An immediate implication of Proposition 2.25 is that all options can be compared among themselves. We recall that we did *not* directly assume that comparisons could be made between all pair of options (an *assumption* which is often criticised as unjustified; see, for example, Fine 1973, p. 221). Instead, we merely assumed that all consequences could be compared among themselves and with the (very simply structured) standard dichotomised options, and that the latter could be compared among themselves.

This completes our elaboration of the axiom system set out in Section 2.3. Starting from the primitive notion of preference, $\leq$, we have shown that quantitative, coherent comparisons of options must proceed *as if* a utility function has been assigned to consequences, probabilities to events and the choice of an option made on the basis of maximising expected utility.

If we *begin* by defining a utility function over $u : \mathcal{C} \to \Re$, this *induces* in turn a preference ordering which is necessarily coherent. Any function can serve as a utility function (subject only to the existence of the expected utility for each option, a problem which does not arise in the case of finite partitions) and the choice is a personal one. In some contexts, however, there are further formal considerations which may delimit the form of function chosen. An important special case is discussed in detail in Section 2.7.

2.6 SEQUENTIAL DECISION PROBLEMS

2.6.1 Complex Decision Problems

Many real decision problems would appear to have a more complex structure than that encapsulated in Definition 2.1. For instance, in the fields of market research and production engineering investigators often consider first whether or not to run a pilot study and only then, in the light of information obtained (or on the basis of initial information if the study is not undertaken), are the major options considered. Such a two-stage process provides a simple example of a *sequential* decision problem, involving successive, interdependent decisions. In this section, we shall demonstrate that complex problems of this kind can be solved with the tools already at our disposal, thus substantiating our claim that the principles of quantitative coherence suffice to provide a prescriptive solution to *any* decision problem.

Before explicitly considering sequential problems, we shall review, using a more detailed notation, some of our earlier developments.

Let $\mathcal{A} = \{a_i, i \in I\}$ be the set of alternative actions we are willing to consider. For each a_i, there is a class $\{E_{ij}, j \in J_i\}$ of exhaustive and mutually exclusive events, which label the possible consequences $\{c_{ij}, j \in J_i\}$ which may result from action a_i. Note that, with this notation, we are merely emphasising the obvious dependence of both the consequences and the events on the action from which they result. If M_0 is our initial state of information and $G > \emptyset$ is additional information obtained subsequently, the main result of the previous section (Proposition 2.25) may be restated as follows.

For behaviour consistent with the principles of quantitative coherence, action a_1 *is to be preferred to action* a_2*, given* M_0 *and* G*, if and only if*

$$\overline{u}(a_1 \,|\, G) > \overline{u}(a_2 \,|\, G),$$

where

$$\overline{u}(a_i \,|\, G) = \sum_{j \in J_i} u(c_{ij}) P(E_{ij} \,|\, a_i, M_0, G),$$

$u(c_{ij})$ *is the value attached to the consequence foreseen if action* a_i *is taken and the event* E_{ij} *occurs, and* $P(E_{ij} \,|\, a_i, M_0, G)$ *is the degree of belief in the occurrence of event* E_{ij}*, conditional on action* a_i *having been taken, and the state of information being* (M_0, G).

We recall that the probability measure used to compute the expected utility is taken to be a representation of the decision-maker's degree of belief conditional on the total information available. By using the extended notation $P(E_{ij} \,|\, a_i, G, M_0)$, rather than the more economical $P(E_j \,|\, G)$ used previously, we are emphasising that (i) the actual events considered may depend on the particular action envisaged, (ii) the information available certainly includes the initial information together with $G > \emptyset$, and (iii) degrees of belief in the occurrence of events such as E_{ij} are understood to be conditional on action a_i having been assumed to be taken, so that the possible influence of the decision-maker on the real world is taken into account.

For any action a_i, it is sometimes convenient to describe the relevant events E_{ij}, $j \in J$, in a sequential form. For example, in considering the relevant events which label the consequences of a surgical intervention for cancer, one may first think of whether the patient will survive the operation and then, conditional on survival, whether or not the tumour will eventually reappear were this particular form of surgery to be performed.

These situations are most easily described diagrammatically using decision trees, such as that shown in Figure 2.4, with as many successive random nodes as necessary. Obviously, this does not represent any formal departure from our previous structure, since the problem can be restated with a single random node where relevant events are defined in terms of appropriate intersections, such as $E_{ij} \cap F_{ijk}$ in the example shown. It is also usually the case, in practice, that it is easier to elicit the relevant degrees of belief conditionally, so that, for example, $P(E_{ij} \cap F_{ijk} \mid a_i, G, M_0)$ would often be best assessed by combining the separately assessed terms $P(F_{ijk} \mid E_{ij}, a_i, G, M_0)$ and $P(E_{ij} \mid a_i, G, M_0)$.

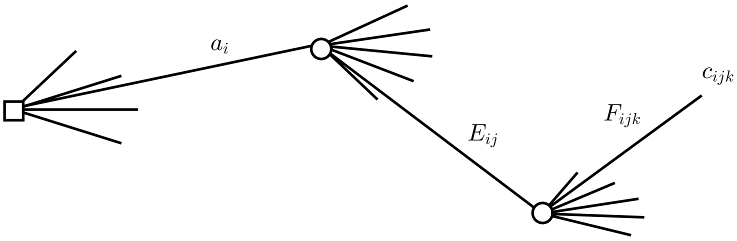

Figure 2.4 *Conditional description of relevant events*

Conditional analysis of this kind is usually necessary in order to understand the structure of complicated situations. Consider, for instance, the problem of placing a bet on the result of a race after which the total amount bet is to be divided up among those correctly guessing the winner. Clearly, if we bet on the favourite we have a higher probability of winning; but, if the favourite wins, many people will have guessed correctly and the prize will be small. It may appear at first sight that this is a decision problem where the utilities involved in an action (the possible prizes to be obtained from a bet) depend on the probabilities of the corresponding uncertain events (the possible winning horses), a possibility *not* contemplated in our structure. A closer analysis reveals, however, that the structure of the problem is similar to that of Figure 2.4. The prize received depends on the bet you place (a_i) the related betting behaviour of other people (E_{ij}) and the outcome of the race (F_{ijk}). It is only natural to assume that our degree of belief in the possible outcomes of the race may be influenced by the betting behaviour of other people. This conditional analysis straightforwardly resolves the initial, apparent complication.

We now turn to considering *sequences* of decision problems. We shall consider situations where, after an action has been taken and its consequences observed, a new decision problem arises, conditional on the new circumstances. For example, when the consequences of a given medical treatment have been observed, a physician has to decide whether to continue the same treatment, or to change to an alternative treatment, or to declare the patient cured.

If a decision problem involves a succession of decision nodes, it is intuitively obvious that the optimal choice at the first decision node depends on the optimal choices at the subsequent decision nodes. In colloquial terms, we typically cannot decide what to do today without thinking first of what we might do tomorrow, and that, of course, will typically depend on the possible consequences of today's actions. In the next section, we consider a technique, *backward induction*, which makes it possible to solve these problems *within* the framework we have already established.

2.6.2 Backward Induction

In any actual decision problem, the number of scenarios which may be contemplated at any given time is necessarily finite. Consequently, and bearing in mind that the analysis is only strictly valid under certain fixed general assumptions and we cannot seriously expect these to remain valid for an indefinitely long period, the number of decision nodes to be considered in any given sequential problem will be assumed to be finite. Thus, we should be able to define a *finite horizon*, after which no further decisions are envisaged in the particular problem formulation. If, at each node, the possibilities are finite in number, the situation may be diagrammatically described by means of a decision tree like that of Figure 2.5.

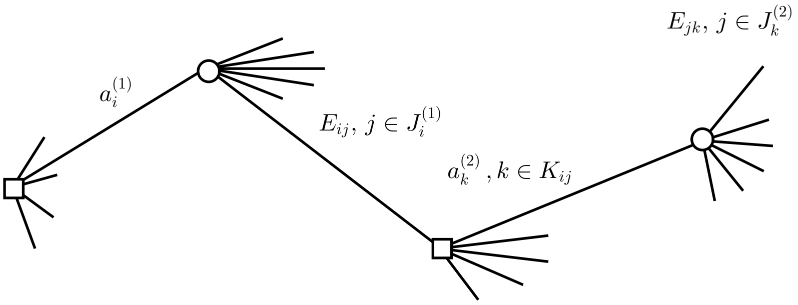
$a_i^{(1)}$
$E_{ij},\ j \in J_i^{(1)}$
$a_k^{(2)}\ , k \in K_{ij}$
$E_{jk},\ j \in J_k^{(2)}$

Figure 2.5 *Decision tree with several decision nodes*

Let n be the number of decision stages considered and let $a^{(m)}$ denote an action being considered at the mth stage. Using the notation for composite options introduced in Section 2.2, all first-stage actions may be compactly described in the form

$$a_i^{(1)} = \left\{ \max_{k \in K_{ij}} a_k^{(2)} \,|\, E_{ij},\ j \in J_i^{(1)} \right\},$$

where $\{E_{ij}, j \in J_i^{(1)}\}$ is the partition of relevant events which corresponds to $a_i^{(1)}$ and the notation "max $a_k^{(2)}$" refers to the *most preferred* of the set of options $\{a_k^{(2)}, k \in K_{ij}\}$ which we would be confronted with were the event E_{ij} to occur. The "maximisation" is naturally to be understood in the sense of our conditional preference ordering among the available second-stage options, given the occurrence of E_{ij}. Indeed, the "consequence" of choosing $a_i^{(1)}$ and having E_{ij} occur is that we are confronted with a set of options $\{a_k^{(2)}, k \in K_{ij}\}$ from which *we can choose* that option which is preferred on the basis of our pattern of preferences at that stage. Similarly, second-stage options may be written in terms of third-stage options, and the process continued until we reach the nth stage, consisting of "ordinary" options defined in terms of the events and consequences to which they may lead. Formally, we have

$$\begin{aligned} a_i^{(m)} &= \left\{ \max_{k \in K_{ij}} a_k^{(m+1)} \,|\, E_{ij},\ j \in J_i^{(m)} \right\}, \quad m = 1, 2, \ldots, n-1, \\ a_i^{(n)} &= \left\{ c_{ij} \,|\, E_{ij}, j \in J_i^{(n)} \right\}. \end{aligned}$$

It is now apparent that sequential decision problems are a special case of the general framework which we have developed.

It follows from Proposition 2.25 that, at each stage m, if G_m is the relevant information available, and $u(.)$ is the (generalised) utility function, we may write

$$a_i^{(m)} \le_{G_m} a_j^{(m)} \iff \overline{u}\left\{ a_i^{(m)} \,|\, G_m \right\} \le \overline{u}\left\{ a_j^{(m)} \,|\, G_m \right\},$$

where

$$\begin{aligned} \overline{u}\left\{ a_i^{(m)} \,|\, G_m \right\} &= \sum_{j \in J_i^{(m)}} \left[\max_{k \in K_{ij}} \overline{u}\left\{ a_k^{(m+1)} \,|\, G_{m+1} \right\} \right] P(E_{ij} \,|\, G_m), \\ \overline{u}\left\{ a_i^{(n)} \,|\, G_n \right\} &= \sum_{j \in J_i^{(n)}} u(c_{ij}) P(E_{ij} \,|\, G_n). \end{aligned}$$

This means that one has to first solve the final (nth) stage, by maximising the appropriate expected utility; then one has to solve the $(n-1)$th stage by maximizing the expected utility conditional on making the optimal choice at the nth stage; and so on, working backwards progressively, until the optimal first stage option has been obtained, a procedure often referred to as *dynamic programming*.

This process of *backward induction* satisfies the requirement that, at any stage of the procedure, the mth, say, the continuation of the procedure must be identical to the optimal procedure starting at the mth stage with information G_m. This requirement is usually known as *Bellman's optimality principle* (Bellman, 1957). As with the "principle" of maximising expected utility, we see that this is not required as a further assumed "principle" in our formulation, but is simply a consequence of the principles of quantitative coherence.

Example 2.4. ***(An optimal stopping problem)***. We now consider a famous problem, which is usually referred to in the literature as the "marriage problem" or the "secretary problem". Suppose that a specified number of objects $n \geq 2$ are to be inspected sequentially, one at a time, in order to select one of them. Suppose further that, at any stage r, $1 \leq r \leq n$, the inspector has the option of either stopping the inspection process, receiving, as a result, the object currently under inspection, or of continuing the inspection process with the next object. No backtracking is permitted and if the inspection process has not terminated before the nth stage the outcome is that the nth object is received. At each stage, r, the only information available to the inspector is the relative rank (1=best, r=worst) of the current object among those inspected so far, and the knowledge that the n objects are being presented in a completely random order.

When should the inspection process be terminated? Intuitively, if the inspector stops too soon there is a good chance that objects more preferred to those seen so far will remain uninspected. However, if the inspection process goes on too long there is a good chance that the overall preferred object will already have been encountered and passed over.

This kind of dilemma is inherent in a variety of practical problems, such as property purchase in a limited seller's market when a bid is required immediately after inspection, or staff appointment in a skill shortage area when a job offer is required immediately after interview. More exotically—and assuming a rather egocentric inspection process, again with no backtracking possibilities—this stopping problem has been suggested as a model for choosing a mate. Potential partners are encountered sequentially; the proverb "marry in haste, repent at leisure" warns against settling down too soon; but such hesitations have to be balanced against painful future realisations of missed golden opportunities.

Less romantically, let c_i, $i = 1, \ldots, n$, denote the possible consequences of the inspection process, with $c_i = i$ if the eventual object chosen has rank i out of all n objects. We shall denote by $u(c_i) = u(i)$, $i = 1, \ldots, n$, the inspector's utility for these consequences.

Now suppose that $r < n$ objects have been inspected and that the relative rank among these of the object under current inspection is x, where $1 \leq x \leq r$. There are two actions available at the rth stage: a_1 = stop, a_2 = continue (where, to simplify notation, we have dropped the superscript, r). The information available at the rth stage is $G_r = (x, r)$; the information available at the $(r + 1)$th stage would be $G_{r+1} = (y, r + 1)$, where y, $1 \leq y \leq r + 1$, is the rank of the next object relative to the $r + 1$ then inspected, all values of y being, of course, equally likely since the n objects are inspected in a random order. If we denote the expected utility of stopping, given G_r, by $\overline{u}_s(x, r)$ and the expected utility of acting optimally, given G_r, by $\overline{u}_0(x,r)$, the general development given above establishes that

$$\overline{u}_0(x,r) = \max\left\{\overline{u}_s(x,r), \frac{1}{r+1}\sum_{y=1}^{r+1}\overline{u}_0(y,r+1)\right\},$$

where

$$\overline{u}_s(x,r) = \sum_{z=x}^{n-r+x} u(z)\frac{\binom{z-1}{x-1}\binom{n-z}{r-x}}{\binom{n}{r}},$$

$$\overline{u}_0(x,n) = \overline{u}_s(x,n) = u(x), \quad x = 1,\ldots,n.$$

Values of $\overline{u}_0(x,r)$ can be found from the final condition and the technique of backwards induction. The optimal procedure is then seen to be:

(i) continue if $\overline{u}_0(x,r) > \overline{u}_s(x,r)$,

(ii) stop if $\overline{u}_0(x,r) = \overline{u}_s(x,r)$.

For illustration, suppose that the inspector's preference ordering corresponds to a "nothing but the best" utility function, defined by $u(1) = 1$, $u(x) = 0$, $x = 2,\ldots,n$. It is then easy to show that

$$\overline{u}_s(1,r) = \frac{r}{n},$$

$$\overline{u}_s(x,r) = 0, \quad x = 2,\ldots,n;$$

thus, if $x > 1$,

$$\overline{u}_0(x,r) > \overline{u}_s(x,r), \quad r = 1,\ldots,n-1.$$

This implies that *inspection should never be terminated if the current object is not the best seen so far*. The decision as to whether to stop if $x = 1$ is determined from the equation

$$\overline{u}_0(x,r) = \max\left\{\frac{r}{n}, \frac{r}{n}\left(\frac{1}{n-1} + \cdots + \frac{1}{r}\right)\right\},$$

which is easily verified by induction. If r^* is the smallest positive integer for which

$$\frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{r^*} \le 1,$$

the optimal procedure is defined as follows:

(i) continue until at least r^* objects have been inspected;

(ii) if the r^*th object is the best so far, stop;

(iii) otherwise, continue until the object under inspection is the best so far, then stop (stopping in any case if the nth stage is reached).

If n is large, approximation of the sum in the above inequality by an integral readily yields the approximation $r^* \approx n/e$. For further details, see DeGroot (1970, Chapter 13), whose account is based closely on Lindley (1961a). For reviews of further, related work on this fascinating problem, see Freeman (1983) and Ferguson (1989).

Applied to the problem of "choosing a mate", and assuming that potential partners are encountered uniformly over time between the ages of 16 and 60, the above analysis suggests delaying a choice until one is at least 32 years old, thereafter ending the search as soon as one encounters someone better than anyone encountered thus far. Readers who are suspicious of putting this into practice have the option, of course, of staying at home and continuing their study of this volume.

Sequential decision problems are now further illustrated by considering the important special case of situations involving an initial choice of experimental design.

2.6.3 Design of Experiments

A simple, very important example of a sequential problem is provided by the situation where we have available a class of experiments, one of which is to be performed in order to provide information for use in a subsequent decision problem. We want to choose the "best" experiment. The structure of this problem, which embraces the topic usually referred to as the problem of *experimental design*, may be diagrammatically described by means of a sequential decision tree such as that shown in Figure 2.6.

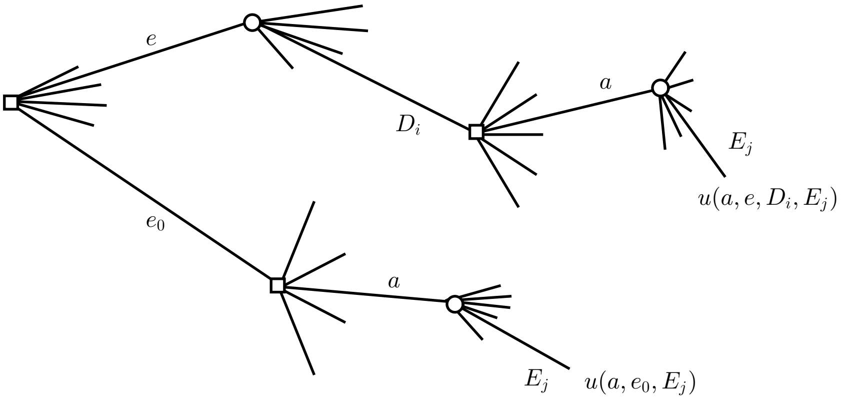

Figure 2.6 *Decision tree for experimental design*

We must first choose an experiment e and, in light of the data D obtained, take an action a, which, were event E to occur, would produce a consequence having utility which, modifying earlier notation in order to be explicit about the elements involved, we denote by $u(a, e, D, E)$. Usually, we also have available the possibility, denoted by e_0 and referred to as the *null experiment*, of directly choosing an action without performing any experiment.

Within the general structure for sequential decision problems developed in the previous section, we note that the possible sets of data obtainable may depend on the particular experiment performed, the set of available actions may depend on the results of the experiment performed, and the sets of consequences and labelling events may depend on the particular combination of experiment and action chosen. However, in our subsequent development we will use a simplified notation which suppresses these possible dependencies in order to centre attention on other, more important, aspects of the problem.

We have seen, in Section 2.6.2, that to solve a sequential decision problem we start at the last stage and work backwards. In this case, the expected utility of option a, given the information available at the stage when the action is to be taken, is

$$\overline{u}(a, e, D_i) = \sum_{j \in J} u(a, e, D_i, E_j) P(E_j \,|\, e, D_i, a).$$

For each pair (e, D_i) we can therefore choose the best possible continuation; namely, that action a_i^* which maximises the expression given above. Thus, the expected utility of the pair (e, D_i) is given by

$$\overline{u}(e, D_i) = \overline{u}(a_i^*, e, D_i) = \max_a \overline{u}(a, e, D_i).$$

We are now in a position to determine the best possible experiment. This is that e which maximises, in the class of available experiments, the unconditional expected utility

$$\overline{u}(e) = \sum_{i \in I} \overline{u}(a_i^*, e, D_i) P(D_i \,|\, e),$$

where $P(D_i \,|\, e)$ denotes the degree of belief attached to the occurrence of data D_i if e were the experiment chosen. On the other hand, the expected utility of performing no experiment and choosing that action a_0^* which maximises the (prior) expected utility is

$$\overline{u}(e_0) = \overline{u}(a_0^*, e_0) = \max_a \sum_{j \in J} u(a, e_0, E_j,) P(E_j \,|\, e_0, a),$$

so that an experiment e is worth performing if and only if $\overline{u}(e) > \overline{u}(e_0)$.

Naturally, $\overline{u}(a, e, D_i), \overline{u}(e, D_i)$ and $\overline{u}(e)$ are different functions defined on different spaces. However, to simplify the notation and without danger of confusion we shall always use $\overline{u}$ to denote an expected utility.

Proposition 2.26. (***Optimal experimental design***). *The optimal action is to perform the experiment e^* if $\overline{u}(e^*) > \overline{u}(e_0)$ and $\overline{u}(e^*) = \max_e \overline{u}(e)$; otherwise, the optimal action is to perform no experiment.*

Proof. This is immediate from Proposition 2.25. $\triangleleft$

It is often interesting to determine the value which additional information might have in the context of a given decision problem.

The expected value of the information provided by new data may be computed as the (posterior) expected difference between the utilities which correspond to optimal actions after and before the data have been obtained.

Definition 2.18. (***The value of additional information***).

(i) *The* ***expected value of the data*** D_i *provided by an experiment* e *is*

$$v(e, D_i) = \sum_{j \in J} \left\{ u(a_i^*, e, D_i, E_j) - u(a_0^*, e_0, E_j) \right\} P(E_j \mid e, D_i, a_i^*);$$

where a_i^*, a_0^* *are, respectively, the optimal actions given* D_i, *and with no data.*

(ii) *the* ***expected value of an experiment*** e *is given by*

$$v(e) = \sum_{i \in I} v(e, D_i) p(D_i \mid e).$$

It is sometimes convenient to have an upper bound for the expected value $v(e)$ of an experiment e. Let us therefore consider the optimal actions which would be available with perfect information, i.e., were we to know the particular event E_j which will eventually occur, and let $a_{(j)}^*$ be the optimal action given E_j, i.e., such that, for all E_j,

$$u(a_{(j)}^*, e_0, E_j) = \max_a u(a, e_0, E_j).$$

Then, given E_j, the loss suffered by choosing any other action a will be

$$u(a_{(j)}^*, e_0, E_j) - u(a, e_0, E_j).$$

For $a = a_0^*$, the optimal action under prior information, this difference will measure, conditional on E_j, the value of perfect information and, under appropriate conditions, its expected value will provide an upper bound for the increase in utility which additional data about the E_j's could be expected to provide.

Definition 2.19. (***Expected value of perfect information***). *The* ***opportunity loss*** *which would be suffered if action* a *were taken and event* E_j *occurred is*

$$l(a, E_j) = \max_{a_i} u(a_i, e_0, E_j) - u(a, e_0, E_j);$$

the expected value of perfect information is then given by

$$v^*(e_0) = \sum_{j \in J} l(a_0^*, E_j) P(E_j \mid a_0^*).$$

It is important to bear in mind that the functions $v(D_i), v(e)$ and the number $v^*(e_0)$, all crucially depend on the (prior) probability distributions $\{(P(E_j \,|\, a), a \in \mathcal{A}\}$ although, for notational convenience, we have not made this dependency explicit.

In many situations, the utility function $u(a, e, D_i, E_j)$ may be thought of as made up of two separate components. One is the (experimental) *cost* of performing e and obtaining D_i; the other is the (terminal) *utility* of directly choosing a and then finding that E_j occurs. Often, the latter component does not actually depend on the preceding e and D_i, so that, assuming additivity of the two components, we may write $u(a, e, D_i, E_j) = u(a, e_0, E_j) - c(e, D_i)$ where $c(e, D_i) \geq 0$. Moreover, the probability distributions over the events are often independent of the action taken. When these conditions apply, we can establish a useful upper bound for the expected value of an experiment in terms of the difference between the expected value of complete data and the expected cost of the experiment itself.

Proposition 2.27. (***Additive decomposition***). *If the utility function has the form*

$$u(a, e, D_i, E_j) = u(a, e_0, E_j) - c(e, D_i),$$

with $c(e, D_i) \geq 0$*, and the probability distributions are such that*

$$p(E_j \,|\, e, D_i, a) = p(E_j \,|\, e, D_i), \quad p(E_j \,|\, e_0, a) = p(E_j \,|\, e_0),$$

then, for any available experiment e*,*

$$v(e) \leq v^*(e_0) - \overline{c}(e),$$

where

$$\overline{c}(e) = \sum_{i \in I} c(e, D_i) P(D_i \,|\, e)$$

is the expected cost of e.

Proof. Using Definitions 2.18 and 2.19, $v(e)$ may be written as

$$\sum_{i \in I} \left[\sum_{j \in J} \Big\{ u(a_i^*, e_0, E_j) - c(e, D_i) - u(a_0^*, e_0, E_j) \Big\} P(E_j \,|\, e, D_i) \right] P(D_i \,|\, e)$$

$$= \sum_{i \in I} \left[\max_a \sum_{j \in J} \Big\{ u(a, e_0, E_j) - u(a_0^*, e_0, E_j) \Big\} P(E_j \,|\, e, D_i) \right] P(D_i \,|\, e) - \overline{c}(e)$$

$$\leq \sum_{i \in I} \sum_{j \in J} \left[\max_a u(a, e_0, E_j) - u(a_0^*, e_0, E_j) \right] P(E_j \cap D_i \,|\, e) - \overline{c}(e)$$

and, hence,

$$\begin{aligned}
v(e) &\leq \Big\{\sum_{j\in J} l(a,E_j)P(E_j\,|\,e_0)\Big\}\Big\{\sum_{i\in I} P(D_i\,|\,E_j,e)\Big\} - \overline{c}(e)\\
&= \Big\{\sum_{j\in J} l(a,E_j)P(E_j\,|\,e_0)\Big\} - \overline{c}(e)\\
&= v^*(e_0) - \overline{c}(e),
\end{aligned}$$

as stated. $\triangleleft$

In Section 2.7, we shall study in more detail the special case of experimental design in situations where data are being collected for the purpose of pure inference, rather than as an input into a directly practical decision problem.

We have shown that the simple decision problem structure introduced in Section 2.2, and the tools developed in Sections 2.3 to 2.5, suffice for the analysis of complex, sequential problems which, at first sight, appear to go beyond that simple structure. In particular, we have seen that the important problem of experimental design can be analysed within the sequential decision problem framework. We shall now use this framework to analyse the very special form of decision problem posed by *statistical inference*, thus establishing the fundamental relevance of these foundational arguments for statistical theory and practice.

2.7 INFERENCE AND INFORMATION

2.7.1 Reporting Beliefs as a Decision Problem

The results on quantitative coherence (Sections 2.2 to 2.5) establish that if we aspire to analyse a given decision problem, $\{\mathcal{E},\mathcal{C},\mathcal{A},\leq\}$, in accordance with the axioms of quantitative coherence, we must represent degrees of belief about uncertain events in the form of a finite probability measure over $\mathcal{E}$ and values for consequences in the form of a utility function over $\mathcal{C}$. Options are then to be compared on the basis of expected utility.

The probability measure represents an individual's beliefs conditional on his or her current state of information. Given the initial state of information described by M_0 and further information in the form of the assumed occurrence of a significant event G, we previously denoted such a measure by $P(.\,|\,G)$. We now wish to specialise our discussion somewhat to the case where G can be thought of as a description of the outcome of an investigation (typically a survey, or an experiment) involving the deliberate collection of data (usually, in numerical form). The event G will then be defined directly in terms of the counts or measurements obtained, either as a precise statement, or involving a description of intervals within which readings lie. To emphasise the fact that G characterises the actual *data* collected, we shall denote the event which describes the new information obtained by D. An individual's degree of belief measure over $\mathcal{E}$ will then be denoted $P(.\,|\,D)$ representing the individual's current beliefs in the light of the data obtained (where, again, we have suppressed, for notational convenience, the explicit dependence on M_0). So far as uncertainty about the events of $\mathcal{E}$ is concerned, $P(.\,|\,D)$ constitutes a complete encapsulation of the information provided by D, given the initial state of information M_0. Moreover, in conjunction with the specification of a utility function, $P(.\,|\,D)$ provides all that is necessary for the calculation of the expected utility of any option and, hence, for the solution of *any decision problem* defined in terms of the frame of reference adopted.

Starting from the decision problem framework, we thus have a formal justification for the main topic of this book; namely, *the study of models and techniques for analysing the ways in which beliefs are modified by data.* However, many eminent writers have argued that basic problems of reporting scientific inferences do not fall within the framework of decision problems as defined in earlier sections:

> Statistical inferences involve the data, a specification of the set of possible populations sampled and a question concerning the true population... Decisions are based on not only the considerations listed for inferences, but also on an assessment of the losses resulting from wrong decisions... (Cox, 1958);
>
> ... a considerable body of doctrine has attempted to explain, or rather to reinterpret these (significance) tests on the basis of quite a different model, namely as means to making decisions in an acceptance procedure. The differences between these two situations seem to the author many and wide, ... (Fisher, 1956/1973).

If views such as these were accepted, they would, of course, undermine our conclusion that *problems concerning uncertainty are to be solved by revising degrees of belief in the light of new data in accordance with Bayes' theorem.* Our main purpose in this section is therefore to demonstrate that *the problem of reporting inferences is essentially a special case of a decision problem.*

By way of preliminary clarification, let us recall from Section 2.1 that we distinguished two, possibly distinct, reasons for trying to think rationally about uncertainty. On the one hand, quoting Ramsey (1926), we noted that, even if an immediate decision problem does not appear to exist, we know that our statements of uncertainty may be used by others in contexts representable within the decision framework. In such situations, our conclusion holds. On the other hand, quoting Lehmann (1959/1986), we noted that the inference, or inference statement, may sometimes be regarded *as an end in itself*, to be judged independently of any "practical" decision problem. It is this case that we wish to consider in more detail in this section, establishing that, indeed, it can be regarded as falling within the general framework of Sections 2.2 to 2.5.

Formalising the first sentence of the remark of Cox, given above, a pure inference problem may be described as one in which we seek to learn which of a set of mutually exclusive "hypotheses" ("theories", "states of nature", or "model parameters") is true. From a strictly realistic viewpoint, there is always, implicitly, a finite set of such hypotheses, say $\{H_j,\, j \in J\}$, although it may be mathematically convenient to work as if this were not the case. We shall regard this set of hypotheses as equivalent to a finite partition of the certain event into events $\{E_j,\, j \in J\}$, having the interpretation $E_j \equiv$ "the hypothesis H_j is true". The actions available to an individual are the various inference statements that might be made about the events $\{E_j,\, j \in J\}$, the latter constituting the uncertain events corresponding to each action. To complete the basic decision problem framework, we need to acknowledge that, corresponding to each *inference statement* and each E_j, there will be a *consequence*; namely, the record of what the individual put forward as an appropriate inference statement, together with what actually turned out to be the case.

If we aspire to quantitative coherence in such a framework, we know that our uncertainty about the $\{E_j,\, j \in J\}$ should be represented by $\{P(E_j \,|\, D),\, j \in J\}$, where $P(.\,|\,D)$ denotes our current degree of belief measure, given data D in addition to the initial information M_0. It is natural, therefore, to regard the set of possible *inference statements* as the class of probability distributions over $\{E_j,\, j \in J\}$ compatible with the information D. The inference reporting problem can thus be viewed as one of choosing a probability distribution to serve as an inference statement. *But there is nothing (so far) in this formulation which leads to the conclusion that the best action is to state one's actual beliefs.* Indeed, we know from our earlier development that options cannot be ordered without an (implicit or explicit) specification of utilities for the consequences. We shall consider this specification and its implications in the following sections. A particular form of utility function for inference statements will be introduced and it will then be seen that the idea of *inference as decision* leads to rather natural interpretations of commonly used information measures in terms of expected utility. In the discussion which follows, we shall only consider the case of finite partitions $\{E_j,\, j \in J\}$. Mathematical extensions will be discussed in Chapter 3.

2.7.2 The Utility of a Probability Distribution

We have argued above that the provision of a statistical inference statement about a class of exclusive and exhaustive "hypotheses" $\{E_j,\, j \in J\}$, conditional on some relevant data D, may be precisely stated as a decision problem, where the set of "hypotheses" $\{E_j,\, j \in J\}$ is a partition consisting of elements of $\mathcal{E}$, and the action space $\mathcal{A}$ relates to the class $\mathcal{Q}$ of conditional probability distributions over $\{E_j,\, j \in J\}$; thus,

$$\mathcal{Q} = \Big\{ \boldsymbol{q} \equiv (q_j,\, j \in J); \quad q_j \geq 0, \quad \textstyle\sum_{j \in J} q_j = 1 \Big\},$$

where q_j is assumed to be the probability which, conditional on the available data D, an individual *reports* as the probability of $E_j \equiv H_j$ being true. The set of consequences $\mathcal{C}$, consists of all pairs $(\boldsymbol{q}, E_j)$, representing the conjunctions of reported beliefs and true hypotheses. The action corresponding to the choice of $\boldsymbol{q}$ is defined as $\{(\boldsymbol{q}, E_j) \,|\, E_j,\ j \in J\}$.

To avoid triviality, we assume that none of the hypotheses is certain and that, without loss of generality, all are compatible with the available data; i.e., that all the E_j's are significant given D, so that (Proposition 2.5) $\emptyset < E_j \cap D < D$ for all $j \in J$. If this were not so, we could simply discard any incompatible hypotheses. It then follows from Proposition 2.17(iii) that each of the personal degrees of belief attached by the individual to the conflicting hypotheses given the data must be strictly positive. Throughout this section, we shall denote by

$$\boldsymbol{p} \equiv (p_j = P(E_j \,|\, D),\, j \in J)\ , \quad p_j > 0, \quad \sum\nolimits_{j \in J} p_j = 1,$$

the probability distribution which describes, conditional again on the available data D, the individual's *actual* beliefs about the alternative "hypotheses".

We emphasise again that, in the structure described so far, there is no logical requirement which forces an individual to *report* the probability distribution $\boldsymbol{p}$ which describes his or her personal *beliefs*, in preference to any other probability distribution $\boldsymbol{q}$ in $\mathcal{Q}$.

We complete the specification of this decision problem by inducing the preference ordering through direct specification of a utility function $u(.)$, which describes the "value" $u(\boldsymbol{q}, E_j)$ of reporting the probability distribution $\boldsymbol{q}$ as the final inferential summary of the investigation, were E_j to turn out to be the true "state of nature". Our next task is to investigate the properties which such a function should possess in order to describe a preference pattern which accords with what a scientific community ought to demand of an inference statement. This special class of utility functions is often referred to as the class of score functions (see also Section 2.8) since the functions describe the possible "scores" to be awarded to the individual as a "prize" for his or her "prediction".

Definition 2.20. (***Score function***). *A score function u for probability distributions $\boldsymbol{q} = \{q_j,\ j \in J\}$ defined over a partition $\{E_j,\ j \in J\}$ is a mapping which assigns a real number $u\{\boldsymbol{q}, E_j\}$ to each pair $(\boldsymbol{q}, E_j)$. This function is said to be* ***smooth*** *if it is continuously differentiable as a function of each q_j.*

It seems natural to assume that score functions should be smooth (in the intuitive sense), since one would wish small changes in the reported distribution to produce only small changes in the obtained score. The mathematical condition imposed is a simple and convenient representation of such smoothness.

We have characterised the problem faced by an individual reporting his or her beliefs about conflicting "hypotheses" as a problem of choice among probability distributions over $\{E_j,\, j \in J\}$, with preferences described by a score function. This is a well specified problem, whose solution, in accordance with our development based on quantitative coherence, is to report that distribution $\boldsymbol{q}$ which maximises the expected utility

$$\sum_{j\in J} u(\boldsymbol{q}, E_j)\; P(E_j \,|\, D).$$

In order to ensure that a coherent individual is also *honest*, we need a form of $u(.)$ which guarantees that the expected utility is maximised if, and only if, $q_j = p_j = P(E_j \,|\, D)$, for each j; otherwise, the individual's best policy could be to report something other than his or her true beliefs. This motivates the following definition:

Definition 2.21. (***Proper score function***). *A score function u is proper if, for each strictly positive probability distribution $\mathbf{p} = \{p_j,\, j \in J\}$ defined over a partition $\{E_j,\, j \in J\}$,*

$$\sup_{\boldsymbol{q}\in\mathcal{Q}} \left\{ \sum_{j\in J} u(\boldsymbol{q}, E_j)p_j \right\} = \sum_{j\in J} u(\boldsymbol{p}, E_j)p_j\,,$$

where the supremum, taken over the class $\mathcal{Q}$ of all probability distributions over $\{E_j,\, j \in J\}$, is attained if, and only if, $\boldsymbol{q} = \mathbf{p}$.

It would seem reasonable that, in a scientific inference context, one should require a score function to be proper. Whether a scientific report presents the inference of a single scientist or a range of inferences, purporting to represent those that might be made by some community of scientists, we should wish to be reassured that any reported inference could be justified as a *genuine* current belief.

Smooth, proper score functions have been successfully used in practice in the following contexts: (i) to determine an appropriate fee to be paid to meteorologists in order to encourage them to report reliable predictions (Murphy and Epstein, 1967); (ii) to score multiple choice examinations so that students are encouraged to assign, over the possible answers, probability distributions which truly describe their beliefs (de Finetti, 1965; Bernardo, 1981b, Section 3.6); (iii) to devise general procedures to elicit personal probabilities and expectations (Savage, 1971); (iv) to select best subsets of variables for prediction purposes in political or medical contexts (Bernardo and Bermúdez, 1985).

The simplest proper score function is the quadratic function (Brier, 1950; de Finetti, 1962) defined as follows.

Definition 2.22. (***Quadratic score function***). *A quadratic score function for probability distributions* $\boldsymbol{q} = \{q_j, j \in J\}$ *defined over a partition* $\{E_j,\, j \in J\}$ *is any function of the form*

$$u\{\boldsymbol{q}, E_j\} = A\left\{2q_j - \sum_{i \in J} q_i^2\right\} + B_j, \quad A > 0,$$

where $\boldsymbol{q} = \{q_j,\, j \in J\}$ *is any probability distribution over* $\{E_j,\, j \in J\}$.

Using the indicator function for E_j, 1_{Ej}, an alternative expression for the quadratic score function is given by

$$u\{\boldsymbol{q}, E_j\} = A\left\{1 - \sum_{i \in J} \left(q_i - 1_{Ej}\right)^2\right\} + B_j, \quad A > 0,$$

which makes explicit the role of a 'penalty' equal to the squared euclidean distance from $\boldsymbol{q}$ to a perfect prediction.

Proposition 2.28. *A quadratic score function is proper.*

Proof. We have to maximise, over $\boldsymbol{q}$, the expected score

$$\sum_{j \in J} u\{\boldsymbol{q}, E_j\}\, p_j = \sum_{j \in J} \left\{A\left(2q_j - \sum_{i \in J} q_i^2\right) + B_j\right\} p_j\,.$$

Taking derivatives with respect to the q_j's and equating them to zero, we have the system of equations $2p_j - 2q_j\{\sum_k p_k\} = 0, j \in J$, and since $\sum_i p_i = 1$, we have $q_j = p_j$ for all j. It is easily checked that this gives a maximum. $\triangleleft$

Note that in the proof of Proposition 2.28 we did *not* need to use the condition $\sum_j q_j = 1$; this is a rather special feature of the quadratic score function.

A further condition is required for score functions in contexts, which we shall refer to as "pure inference problems", where the value of a distribution, $\boldsymbol{q}$, is only to be assessed in terms of the probability it assigned to the actual outcome.

Definition 2.23. (***Local score function***). *A score function* u *is local if, for each element* $\boldsymbol{q} = \{q_j,\, j \in J\}$ *of the class* $\mathcal{Q}$ *of probability distributions defined over a partition* $\{E_j,\, j \in J\}$, *there exist functions* $\{u_j(.),\, j \in J\}$ *such that* $u\{\boldsymbol{q}, E_j\} = u_j(q_j)$.

It is intuitively clear that the preferences of an individual scientist faced with a pure inference problem should correspond to the ordering induced by a local score function. The reason for this is that, by definition, in a "pure" inference problem we are solely concerned with "the truth". It is therefore natural that if E_j, say, turns out to be true, the individual scientist should be assessed (i.e., scored) only on the basis of his or her reported judgement about the plausibility of E_j.

> This can be contrasted with the forms of "score" function that would typically be appropriate in more directly practical contexts. In stock control, for example, probability judgements about demand would usually be assessed in the light of the relative seriousness of under- or over-stocking, rather than by just concentrating on the belief previously attached to what turned out to be the actual level of demand.

Note that, in Definition 2.23, the functional form $u_j(p_j)$ of the dependence of the score on the probability attached to the true E_j is allowed to vary with the particular E_j considered. By permitting different $u_j(.)$'s for each E_j, we allow for the possibility that "bad predictions" regarding some "truths" may be judged more harshly than others.

The situation described by a local score function is, of course, an idealised, limit situation, but one which seems, at least approximately, appropriate in reporting pure scientific research. In addition, later in this section we shall see that certain well-known criteria for choosing among experimental designs are optimal if, and only if, preferences are described by a smooth, proper, local score function.

Proposition 2.29. (***Characterisation of proper local score functions***). *If u is a smooth, proper, local score function for probability distributions $\boldsymbol{q} = \{q_j, j \in J\}$ defined over a partition $\{E_j,\ j \in J\}$ which contains more than two elements, then it must be of the form $u\{\boldsymbol{q}, E_j\} = A \log q_j + B_j$, where $A > 0$ and the B_j's are arbitrary constants.*

Proof. Since $u(.)$ is local and proper, then for some $\{u_j(.), j \in J\}$, we must have

$$\sup_{\boldsymbol{q}} \sum_{j\in J} u(\boldsymbol{q}, E_j)\, p_j = \sup_{\boldsymbol{q}} \sum_{j\in J} u_j(q_j)\, p_j = \sum_{j\in J} u_j(p_j)\, p_j,$$

where $p_j > 0,\ \sum_j p_j = 1$ and the supremum is taken over the class of probability distributions $\boldsymbol{q} = (q_j, j \in J),\ q_j \geq 0,\ \sum_j q_j = 1$.

Writing $\boldsymbol{p} = \{p_1, p_2, \ldots\}$ and $\boldsymbol{q} = \{q_1, q_2, \ldots\}$, with

$$p_1 = 1 - \sum_{j>1} p_j, \quad q_1 = 1 - \sum_{j>1} q_j,$$

we seek $\{u_j(.)\,, j \in J\}$, giving an extremal of

$$F\{q_2, q_3, \ldots\} = \left(1 - \sum_{j>1} p_j\right) u_1 \left(1 - \sum_{j>1} q_j\right) + \sum_{j>1} p_j u_j(q_j)\,,$$

For $\{q_2, q_3, \ldots\}$ to make F stationary it is necessary (see e.g. Jeffreys and Jeffreys, 1946, p. 315) that

$$\frac{\partial}{\partial\alpha} F\{q_2 + \alpha\varepsilon_2, q_3 + \alpha\varepsilon_3, \ldots\}\Big|_{\alpha=0} = 0$$

for any $\varepsilon = \{\varepsilon_2, \varepsilon_3, \ldots\}$ such that all the ε_j are sufficiently small. Calculating this derivative, the condition is seen to reduce to

$$\sum_{j>1} \left\{ \left(1 - \sum_{i>1} p_i\right) u_1' \left(1 - \sum_{j>1} q_j\right) - p_j u_j'(q_j) \right\} \varepsilon_j = 0$$

for all ε_j's sufficiently small, where u' stands for the derivative of u. Moreover, since u is proper, $\{p_2, p_3, \ldots\}$ must be an extremal of F and thus we have the system of equations

$$p_j \, u_j'(p_j) = \left(1 - \sum_{i>1} p_i\right) u_1' \left(1 - \sum_{i>1} p_i\right), \quad j = 2, 3, \ldots$$

so that all the functions $u_j, j = 1, 2, \ldots$ satisfy the same functional equation, namely

$$p_j \, u_j'(p_j) = p_1 \, u_1'(p_1), \quad j = 2, 3, \ldots,$$

for all $\{p_2, p_3, \ldots\}$ and, hence,

$$p \, u_j'(p) = A, \quad 0 < p \leq 1, \qquad \text{for all } j = 1, 2, \ldots$$

so that $u_j(p) = A \log p + B_j$. The condition $A \geq 0$ suffices to guarantee that the extremal found is indeed a maximum. $\triangleleft$

Definition 2.24. (***Logarithmic score function***). *A logarithmic score function for strictly positive probability distributions $\boldsymbol{q} = \{q_j, j \in J\}$ defined over a partition $\{E_j, \, j \in J\}$ is any function of the form*

$$u\{\boldsymbol{q}, E_j\} = A \log \, q_j + B_j, \quad A > 0.$$

If the partition $\{E_j, j \in J\}$ only contains two elements, so that the partition is simply $\{H, H^c\}$, the locality condition is, of course, vacuous. In this case, $u\{\boldsymbol{q}, E_j\} = u\{(q_1, 1 - q_1), \, 1_H\} = f(q_1, 1_H)$, say, where 1_H is the indicator function for H, and the score function only depends on the probability q_1 attached to H, whether or not H occurs.

For $u\{(q_1, 1 - q_1), \, 1_H\}$ to be proper we must have

$$\sup_{q_1 \in [0,1]} \{p_1 \, f(q_1, 1) + (1 - p_1) \, f(q_1, 0)\} = p_1 \, f(p_1, 1) + (1 - p_1) \, f(p_1, 0)$$

so that, if the score function is smooth, then f must satisfy the functional equation

$$x \, f'(x, 1) + (1 - x) \, f'(x, 0) = 0.$$

The logarithmic function $f(x,1) = A\log x + B_1$, $f(x,0) = A\log(1-x) + B_2$ is then just one of the many possible solutions (see Good, 1952).

We have assumed that the probability distributions to be considered as options assign strictly positive q_j to each E_j. This means that, given any particular $\boldsymbol{q} \in \mathcal{Q}$, we have no problem in calculating the expected utility arising from the logarithmic score function. It is worth noting, however, that since we place no (strictly positive) lower bound on the possible q_j, we have an example of an unbounded decision problem; i.e., a decision problem without extreme consequences.

2.7.3 Approximation and Discrepancy

We have argued that the optimal solution to an inference reporting problem (either for an individual, or for each of several individuals) is to state the appropriate actual beliefs, $\boldsymbol{p}$, say. From a technical point of view, however, particularly within the mathematical extensions to be considered in Chapter 3, the precise computation of $\boldsymbol{p}$ may be difficult and we may choose instead to report an approximation to our beliefs, $\boldsymbol{q}$, say, on the grounds that $\boldsymbol{q}$ is "close" to $\boldsymbol{p}$, but much easier to calculate. The justification of such a procedure requires a study of the notion of "closeness" between two distributions.

Proposition 2.30. (***Expected loss in probability reporting***). *If preferences are described by a logarithmic score function, the expected loss of utility in reporting a probability distribution $\boldsymbol{q} = \{q_j,\, j \in J\}$ defined over a partition $\{E_j,\, j \in J\}$, rather than the distribution $\boldsymbol{p} = \{p_j,\, j \in J\}$ representing actual beliefs, is given by*

$$\delta\{\boldsymbol{q} \,|\, \boldsymbol{p}\} = A \sum_{j \in J} p_j \log\,(p_j/q_j), \quad A > 0.$$

Moreover, $\delta\{\boldsymbol{q} \,|\, \boldsymbol{p}\} \geq 0$ with equality if and only if $\boldsymbol{q} = \boldsymbol{p}$.

Proof. Using Definition 2.24, the expected utility of reporting $\boldsymbol{q}$ when $\boldsymbol{p}$ is the actual distribution of beliefs is $\overline{u}(\boldsymbol{q}) = \sum_j \{A\log q_j + B_j\}p_j$, and thus

$$\begin{aligned}\delta\{\boldsymbol{q} \,|\, \boldsymbol{p}\} &= \overline{u}(\boldsymbol{p}) - \overline{u}(\boldsymbol{q}) \\ &= \sum_{j \in J} \big\{(A\log p_j + B_j) - (A\log q_j + B_j)\big\}\, p_j = A\sum_{j \in J} p_j \log\frac{p_j}{q_j}\,\cdot\end{aligned}$$

The final statement in the theorem is a consequence of Proposition 2.29 since, because the logarithmic score function is proper, the expected utility of reporting $\boldsymbol{q}$ is maximised if, and only if, $\boldsymbol{q} = \boldsymbol{p}$, so that $\overline{u}(\boldsymbol{p}) \geq \overline{u}(\boldsymbol{q})$, with equality if, and only if, $\boldsymbol{p} = \boldsymbol{q}$. An immediate direct proof is obtained using the fact that for all $x > 0$, $\log x \leq x - 1$ with equality if, and only if, $x = 1$. Indeed, we then have

$$\begin{aligned} -\delta\{\boldsymbol{q} \,|\, \boldsymbol{p}\} &= \sum_{j \in J} p_j \log \frac{q_j}{p_j} \\ &\leq \sum_{j} p_j\{(q_j/p_j) - 1\} = \sum_{j \in J} q_j - \sum_{j \in J} p_j = 1 - 1 = 0, \end{aligned}$$

with equality if, and only if, $q_j = p_j$ for all j. $\triangleleft$

The quantity $\delta\{\boldsymbol{q} \,|\, \boldsymbol{p}\}$, which arises here as a difference between two expected utilities, was introduced by Kullback and Leibler (1951) as an *ad hoc* measure of (directed) *divergence* between two probability distributions.

Combining Propositions 2.29 and 2.30, it is clear that an individual with preferences approximately described by a proper local score function should beware of approximating by zero. This reflects the fact that the "tails" of the distribution are, generally speaking, extremely important in pure inference problems. This is in contrast to many practical decision problems where the form of the utility function often makes the solution robust with respect to changes in the "tails" of the distribution assumed.

Proposition 2.30 suggests a natural, general measure of "lack of fit", or *discrepancy*, between a distribution and an approximation, when preferences are described by a logarithmic score function.

Definition 2.25. (***Discrepancy of an approximation***). *The discrepancy between a strictly positive probability distribution $\boldsymbol{p} = \{p_j,\, j \in J\}$ over a partition $\{E_j,\, j \in J\}$ and an approximation $\hat{\boldsymbol{p}} = \{\hat{p}_j,\, j \in J\}$ is defined by*

$$\delta\{\hat{\boldsymbol{p}} \,|\, \boldsymbol{p}\} = \sum_{j \in J} p_j \log \frac{p_j}{\hat{p}_j} \quad \cdot$$

Example 2.5. ***(Poisson approximation to a binomial distribution)***. The behaviour of $\delta\{\hat{\boldsymbol{p}} \,|\, \boldsymbol{p}\}$ is well illustrated by a familiar, elementary example. Consider the binomial distribution

$$\begin{aligned} p_j &= \binom{n}{j} \theta^j (1-\theta)^{n-j}, \quad j = 0, 1, \ldots, n\,, \\ &= 0\,, \text{ otherwise} \end{aligned}$$

and let

$$\hat{p}_j = \exp\{-n\theta\} \frac{(n\theta)^j}{j!}, \quad j = 0, 1, \ldots$$

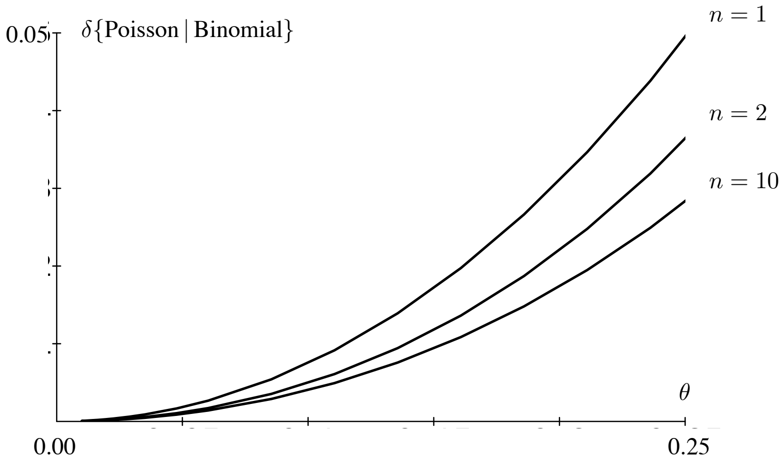

Figure 2.7 *Discrepancy between a binomial distribution and its Poisson approximation (logarithms to base 2).*

be its Poisson approximation. It is apparent from Figure 2.7 that $\delta\{\hat{\boldsymbol{p}} \mid \boldsymbol{p}\}$ decreases as either n increases or θ decreases, or both, and that the second factor is far more important than the first. However, it follows from our previous discussion that it would not be a good idea to reverse the roles and try to approximate a Poisson distribution by a binomial distribution.

When, as in Figure 2.7, logarithms to base 2 are used, the utility and discrepancy are measured on the well-known scale of bits of information (or *entropy*), which can be interpreted in terms of the expected number of yes-no questions required to identify the true event in the partition (see, for example, de Finetti, 1970/1974, p. 103, or Renyi, 1962/1970, p. 564).

Clearly, Definition 2.25 provides a systematic approach to approximation in pure inference contexts. The best *approximation* within a given family will be that which *minimises the discrepancy*.

2.7.4 Information

In Section 2.4.2, we showed that, for quantitative coherence, any new information D should be incorporated into the analysis by updating beliefs via Bayes' theorem, so that the initial representation of beliefs $P(.)$ is updated to the conditional probability measure $P(. \mid D)$. In Section 2.7.2, we showed that, within the context of the pure inference reporting problem, utility is defined in terms of the logarithmic score function.

Proposition 2.31. (***Expected utility of data***). *If preferences are described by a logarithmic score function for the class of probability distributions defined over a partition $\{E_j,\, j \in J\}$, then the expected increase in utility provided by data D, when the initial probability distribution $\{P(E_j), j \in J\}$ is strictly positive, is given by*

$$A \sum_{j \in J} P(E_j \,|\, D) \log \frac{P(E_j \,|\, D)}{P(E_j)}\,,$$

where $A > 0$ is arbitrary, and $\{P(E_j \,|\, D), j \in J\}$ is the conditional probability distribution, given D. Moreover, this expected increase in utility is non-negative and is zero if, and only if, $P(E_j \,|\, D) = P(E_j)$ for all j.

Proof. By Definition 2.24, the utilities of reporting $P(.)$ or $P(.\,|\,D)$, were E_j known to be true, would be $A \log P(E_j) + B_j$ and $A \log P(E_j \,|\, D) + B_j$, respectively. Thus, conditional on D, the expected increase in utility provided by D is given by

$$\sum_{j \in J} \{(A \log P(E_j \,|\, D) + B_j) - (A \log P(E_j) + B_j)\} P(E_j \,|\, D)$$

$$= A \sum_{j \in J} P(E_j \,|\, D) \log \frac{P(E_j \,|\, D)}{P(E_j)}\,,$$

which, by Proposition 2.30, is non-negative and is zero if and only if, for all j, $P(E_j \,|\, D) = P(E_j)$. ◁

In the context of pure inference problems, we shall find it convenient to underline the fact that, because of the use of the logarithmic score function, utility assumes a special form and establishes a link between utility theory and classical information theory. This motivates Definitions 2.26 and 2.27.

Definition 2.26. (***Information from data***). *The amount of information about a partition $\{E_j,\, j \in J\}$ provided by the data D, when the initial distribution over $\{E_j,\, j \in J\}$ is $\boldsymbol{p}_0 = \{P(E_j), j \in J\}$, is defined to be*

$$I(D \,|\, \boldsymbol{p}_0) = \sum_{j \in J} P(E_j \,|\, D) \log \frac{P(E_j \,|\, D)}{P(E_j)}\,,$$

where $\{P(E_j \,|\, D),\, j \in J\}$ is the conditional probability distribution given the data D.

It follows from Definition 2.26 that the amount of information provided by data D is equal to $\delta(\boldsymbol{p}_0 \,|\, \boldsymbol{p}_D)$, the discrepancy measure if $\boldsymbol{p}_0 = \{P(E_j), j \in J\}$ is considered as an approximation to $\boldsymbol{p}_D = \{P(E_j \,|\, D), j \in J\}$. Another interesting interpretation of $I(D \,|\, \boldsymbol{p}_0)$ arises from the following analysis. Conditional on E_j, $\log P(E_j)$ and $\log P(E_j \,|\, D)$ measure, respectively, how good the initial and the conditional distributions are in "predicting" the "true hypothesis" $E_j = H_j$, so that $\log P(E_j \,|\, D) - \log P(E_j)$ is a measure of the value of D, were E_j known to be true; $I(D \,|\, \boldsymbol{p}_0)$ is simply the expected value of that difference calculated with respect to $\boldsymbol{p}_D$.

It should be clear from the preceding discussion that $I(D \,|\, \boldsymbol{p}_0)$ measures indirectly the information provided by the data in terms of the changes produced in the probability distribution of interest. The amount of information is thus seen to be a relative measure, which obviously depends on the initial distribution. Attempts to define absolute measures of information have systematically failed to produce concepts of lasting value.

In the finite case, the *entropy* of the distribution $\boldsymbol{p} = \{p_1, \ldots, p_n\}$, defined by

$$H\{\boldsymbol{p}\} = -\sum_{j=1}^{n} p_j \log p_j,$$

has been proposed and widely accepted as an absolute measure of uncertainty. The recognised fact that its apparently natural extension to the continuous case does not make sense (if only because it is heavily dependent on the particular parametrisation used) should, however, have raised doubts about the universality of this concept. The fact that, in the finite case, $H\{\boldsymbol{p}\}$ as a measure of uncertainty (and $-H\{\boldsymbol{p}\}$ as a measure of "absolute" information) seems to work correctly is explained (from our perspective) by the fact that

$$\sum_{j=1}^{n} p_j \log \frac{p_j}{n^{-1}} = \log n - H\{\boldsymbol{p}\},$$

so that, in terms of the above discussion, $-H\{\boldsymbol{p}\}$ may be interpreted, apart from an unimportant additive constant, as the amount of information which is necessary to obtain $\boldsymbol{p} = \{p_1, \ldots, p_n\}$ from an *initial discrete uniform distribution* (see Section 3.2.2), which acts as an "origin" or "reference" measure of uncertainty. As we shall see in detail later, the problem of extending the entropy concept to continuous distributions is closely related to that of defining an "origin" or "reference" measure of uncertainty in the continuous case, a role unambiguously played by the uniform distribution in the finite case. For detailed discussion of $H\{\boldsymbol{p}\}$ and other proposed entropy measures, see Renyi (1961).

We shall on occasion wish to consider the idea of the amount of information which may be expected from an experiment e, the expectation being calculated before the results of the experiment are actually available.

Definition 2.27. (***Expected information from an experiment***). *The expected information to be provided by an experiment* e *about a partition* $\{E_j,\, j \in J\}$, *when the initial distribution over* $\{E_j,\, j \in J\}$ *is* $\boldsymbol{p}_0 = \{P(E_j),\, j \in J\}$, *is given by*

$$I(e \,|\, \boldsymbol{p}_0) = \sum_{i \in I} I(D_i \,|\, \boldsymbol{p}_0)\, P(D_i),$$

where the possible results of the experiment e, $\{D_i,\, i \in I\}$, *occur with probabilities* $\{P(D_i), i \in I\}$.

Proposition 2.32. *An alternative expression for the expected information is*

$$I(e \,|\, \boldsymbol{p}_0) = \sum_{i \in I} \sum_{j \in J} P(E_j \cap D_i) \log \frac{P(E_j \cap D_i)}{P(E_j)P(D_i)}\,,$$

where $P(E_j \cap D_i) = P(D_i)\; P(E_j \,|\, D_i)$, *and* $\{P(E_j \,|\, D_i), j \in J\}$ *is the conditional distribution, given the occurrence of* D_i, *corresponding to the initial distribution* $\boldsymbol{p}_0 = \{P(E_j), j \in J\}$. *Moreover,* $I(e \,|\, \boldsymbol{p}_0) \geq 0$, *with equality if and only if, for all* E_i *and* D_j, $P(E_j \cap D_i) = P(E_j)P(D_i)$.

Proof. Let $q_i = P(D_i)$, $p_j = P(E_j)$ and $p_{ji} = P(E_j \,|\, D_i)$. Then, by Definition 2.27,

$$I(e \,|\, \boldsymbol{p}_0) = \sum_{i \in I} \left\{ \sum_{j \in J} p_{ji} \log \frac{p_{ji}}{p_j} \right\} q_i \;=\; \sum_{i \in I} \sum_{j \in J} p_{ji} q_i \log \frac{p_{ji} q_i}{p_j q_i}$$

and the result now follows from the fact that, by Bayes' theorem,

$$P(E_j \cap D_i) = P(E_j \,|\, D_i) P(D_i) = p_{ji} q_i.$$

Since, by Proposition 2.31, $I(D_i \,|\, \boldsymbol{p}_0) \geq 0$ with equality iff, $P(E_j \,|\, D_i) = P(E_j)$, it follows from Definition 2.27 that $I(e \,|\, \boldsymbol{p}_0) \geq 0$ with equality if, and only if, for all E_j and D_i, $P(E_j \cap D_i) = P(E_j)P(D_i)$. $\triangleleft$

The expression for $I(e \,|\, \boldsymbol{p}_0)$ given by Proposition 2.32 is Shannon's (1948) measure of expected information. We have thus found, in a decision theoretical framework, a natural interpretation of this famous measure of expected information: *Shannon's expected information is the expected utility provided by an experiment in a pure inference context, when an individual's preferences are described by a smooth, proper, local score function.*

In conclusion, we have suggested that the problem of reporting inferences can be viewed as a particular decision problem and thus should be analysed within the framework of decision theory. We have established that, with a natural characterisation of an individual's utility function when faced with a pure inference problem, preferences should be described by a logarithmic score function. We have also seen that, within this framework, discrepancy and amount of information are naturally defined in terms of expected loss of utility and expected increase in utility, respectively, and that maximising expected Shannon information is a particular instance of maximising expected utility. We shall see in Section 3.4 how these results, established here for finite partitions, extend straightforwardly to the continuous case.

2.8 DISCUSSION AND FURTHER REFERENCES

2.8.1 Operational Definitions

In everyday conversation, the way in which we use language is typically rather informal and unselfconscious, and we tolerate each other's ambiguities and vacuities for the most part, occasionally seeking an *ad hoc* clarification of a particular statement or idea if the context seems to justify the effort required in trying to be a little more precise. (For a detailed account of the ambiguities which plague qualitative probability expressions in English, see Mosteller and Youtz, 1990.)

In the context of scientific and philosophical discourse, however, there is a paramount need for statements which are meaningful and unambiguous. The everyday, tolerant, *ad hoc* response will therefore no longer suffice. More rigorous habits of thought are required, and we need to be selfconsciously aware of the precautions and procedures to be adopted if we are to arrive at statements which make sense.

A prerequisite for "making sense" is that the fundamental concepts which provide the substantive content of our statements should themselves be defined in an essentially unambiguous manner. We are thus driven to seek for definitions of fundamental notions which can be reduced ultimately to the touchstone of actual or potential personal experience, rather than remaining at the level of mere words or phrases.

This kind of approach to definitions is closely related to the philosophy of *pragmatism*, as formulated in the second half of the nineteenth century by Peirce, who insisted that clarity in thinking about concepts could only be achieved by concentrating attention on the conceivable practical effects associated with a concept, or the practical consequence of adopting one form of definition rather than another. In Peirce (1878), this point of view was summarised as follows:

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.

In some respects, however, this position is not entirely satisfactory in that it fails to go far enough in elaborating what is to be understood by the term "practical". This crucial elaboration was provided by Bridgman (1927) in a book entitled *The Logic of Modern Physics*, where the key idea of an *operational* definition is introduced and illustrated by considering the concept of "length":

> . . . what do we mean by the length of an object? We evidently know what we mean by length if we can tell what the length of any and every object is and for the physicist nothing more is required. To find the length of an object, we have to perform certain physical operations. The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as, and nothing more, than the set of operations by which length is determined. In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations. If the concept is physical, . . . the operations are actual physical measurements . . . ; or if the concept is mental, . . . the operations are mental operations. . .

Throughout this work, we shall seek to adhere to the operational approach to defining concepts in order to arrive at meaningful and unambiguous statements in the context of representing beliefs and taking actions in situations of uncertainty. Indeed, we have stressed this aspect of our thinking in Sections 2.1 to 2.7, where we made the practical, operational idea of preference between options the fundamental starting point and touchstone for all other definitions.

We also noted the inevitable element of idealisation, or approximation, implicit in the operational approach to our concepts, and we remarked on this at several points in Section 2.3. Since many critics of the personalistic Bayesian viewpoint claim to find great difficulty with this feature of the approach, often suggesting that it undermines the entire theory, it is worth noting Bridgman's very explicit recognition that *all* experience is subject to error and that all we can do is to take sufficient precautions when specifying sets of operations to ensure that remaining unspecified variations in procedure have negligible effects on the results of interest. This is well illustrated by Bridgman's account of the operational concept of length and its attendant idealisations and approximations:

> . . .we take a measuring rod, lay it on the object so that one of its ends coincides with one end of the object, mark on the object the position of the rod, then move the rod along in a straight line extension of its previous position until the first end coincides with the previous position of the second end, repeat this process as often as we can, and call the length the total number of times the rod was applied. This procedure, apparently so simple, is in practice exceedingly complicated, and doubtless a full description of all the precautions that must be taken would fill a large treatise. We must, for example, be sure that the temperature of the rod is the standard temperature at which its length is defined, or else we must make a

> correction for it; or we must correct for the gravitational distortion of the rod if we measure a vertical length; or we must be sure that the rod is not a magnet or is not subject to electrical forces . . . we must go further and specify all the details by which the rod is moved from one position to the next on the object, its precise path through space and its velocity and acceleration in getting from one position to another. Practically, of course, precautions such as these are not taken, but the justification is in our experience that variations of procedure of this kind are without effect on the final result. . .

This pragmatic recognition that there are inevitable limitations in any concrete application of a set of operational procedures is precisely the spirit of our discussion of Axioms 4 and 5 in Section 2.3. In practical terms, we have to stop somewhere, even though, in principle, we could indefinitely refine our measurement operations. What matters is to be able to achieve sufficient accuracy to avoid unacceptable distortion in any analysis of interest.

2.8.2 Quantitative Coherence Theories

In a comprehensive review of normative decision theories leading to the expected utility criterion, Fishburn (1981) lists over thirty different axiomatic formulations of the principles of coherence, reflecting a variety of responses to the underlying conflict between axiomatic simplicity and structural flexibility in the representation of decision problems. Fishburn sums up the dilemma as follows:

> On the one hand, we would like our axioms to be simple, interpretable, intuitively clear, and capable of convincing others that they are appealing criteria of coherency and consistency in decision making under uncertainty, but to do this it seems essential to invoke strong structural conditions. On the other hand, we would like our theory to adhere to the loose structures that often arise in realistic decision situations, but if this is done then we will be faced with fairly complicated axioms that accommodate these loose structures.

In addition, we should like the definitions of the basic concepts of probability and utility to have strong and direct links with practical assessment procedures, in conformity with the operational philosophy outlined above.

With these considerations in mind, our purpose here is to provide a brief historical review of the foundational writings which seem to us the most significant. This will serve in part to acknowledge our general intellectual indebtedness and orientation, and in part to explain and further motivate our own particular choice of axiom system.

The earliest axiomatic approach to the problem of decision making under uncertainty is that of Ramsey (1926), who presented the outline of a formal system. The key postulate in Ramsey's theory is the existence of a so-called *ethically neutral* event E, say, which, expressed in terms of our notation for options, has the property that $\{c_1 \,|\, E, c_2 \,|\, E^c\} \sim \{c_1 \,|\, E^c, c_2 \,|\, E\}$, for any consequences c_1, c_2. It is then rather natural to define the degree of belief in such an event to be $1/2$ and, from this quantitative basis, it is straightforward to construct an operational measure of utility for consequences. This, in turn, is used to extend the definition of degree of belief to general events by means of an expected utility model.

From a conceptual point of view, Ramsey's theory seems to us, as indeed it has to many other writers, a revolutionary landmark in the history of ideas. From a mathematical point of view, however, the treatment is rather incomplete and it was not until 1954, with the publication of Savage's (1954) book *The Foundations of Statistics* that the first complete formal theory appeared. No mathematical completion of Ramsey's theory seems to have been published, but a closely related development can be found in Pfanzagl (1967, 1968).

Savage's major innovation in structuring decision problems is to define what he calls acts (options, in our terminology) as functions from the set of uncertain possible outcomes into the set of consequences. His key coherence assumption is then that of a complete, transitive order relation among acts and this is used to define qualitative probabilities. These are extended into quantitative probabilities by means of a "continuously divisible" assumption about events. Utilities are subsequently introduced using ideas similar to those of von Neumann and Morgenstern (1944/1953), who had, ten years earlier, presented an axiom system for utility alone, assuming the prior existence of probabilities.

The Savage axiom system is a great historical achievement and provides the first formal justification of the personalistic approach to probability and decision making; for a modern appraisal see Shafer (1986) and lively ensuing discussion. See, also, Hens (1992). Of course, many variations on an axiomatic theme are possible and other Savage-type axiom systems have been developed since by Stigum (1972), Roberts (1974), Fishburn (1975) and Narens (1976). Suppes (1956) presented a system which combined elements of Savage's and Ramsey's approaches. See, also, Suppes (1960, 1974) and Savage (1970). There are, however, two major difficulties with Savage's approach, which impose severe limitations on the range of applicability of the theory.

The first of these difficulties stems from the "continuously divisible" assumption about events, which Savage uses as the basis for proceeding from qualitative to quantitative concepts. Such an assumption imposes severe constraints on the allowable forms of structure for the set of uncertain outcomes: in fact, it even prevents the theory from being directly applicable to situations involving a finite or countably infinite set of possible outcomes.

One way of avoiding this embarrassing structural limitation is to introduce a quantitative element into the system by a device like that of Ramsey's ethically neutral event. This is directly defined to have probability $1/2$ and thus enables Ramsey to get the quantitative ball rolling without imposing undue constraints on the structure. All he requires is that (at least) one such event be included in the representation of the uncertain outcomes. In fact, a generalisation of Ramsey's idea re-emerges in the form of canonical lotteries, introduced by Anscombe and Aumann (1963) for defining degrees of belief, and by Pratt, Raiffa and Schlaifer (1964, 1965) as a basis for simultaneously quantifying personal degrees of belief and utilities in a direct and intuitive manner.

The basic idea is essentially that of a standard measuring device, in some sense external to the real-world events and options of interest. It seems to us that this idea ties in perfectly with the kind of operational considerations described above, and the standard events and options that we introduced in Section 2.3 play this fundamental operational role in our own system. Other systems using standard measuring devices (sometimes referred to as external scaling devices) are those of Fishburn (1967b, 1969) and Balch and Fishburn (1974). A theory which, like ours, combines a standard measuring device with a fundamental notion of *conditional* preference is that of Luce and Krantz (1971).

The second major difficulty with Savage's theory, and one that also exists in many other theories (see Table I in Fishburn, 1981), is that the Savage axioms imply the boundedness of utility functions (an implication of which Savage was apparently unaware when he wrote *The Foundations of Statistics*, but which was subsequently proved by Fishburn, 1970). The theory does not therefore justify the use of many mathematically convenient and widely used utility functions; for example, those implicit in forms such as "quadratic loss" and "logarithmic score".

We take the view, already hinted at in our brief discussion of medical and monetary consequences in Section 2.5, that it is often conceptually and mathematically convenient to be able to use structural representations going beyond what we perceive to be the essentially finitistic and bounded characteristics of real-world problems. And yet, in presenting the basic quantitative coherence axioms it is important not to confuse the primary definitions and coherence principles with the secondary issues of the precise forms of the various sets involved. For this reason, we have so far always taken options to be defined by finite partitions; indeed, within this simple structure, we hope that the essence of the quantitative coherence theory has already been clearly communicated, uncomplicated by structural complexities. Motivated by considerations of mathematical convenience, however, we shall, in Chapter 3, relax the constraint imposed on the form of the action space. We shall then arrive at a sufficiently general setting for all our subsequent developments and applications.

2.8.3 Related Theories

Our previous discussion centred on complete axiomatic approaches to decision problems, involving a unified development of both probability and utility concepts. In our view, a unified treatment of the two concepts is inescapable if operational considerations are to be taken seriously. However, there have been a number of attempted developments of probability ideas separate from utility considerations, as well as separate developments of utility ideas presupposing the existence of probabilities. In addition, there is a considerable literature on information-theoretic ideas closely related to those of Section 2.7. In this section, we shall provide a summary overview of a number of these related theories, grouped under the following subheadings: (i) *Monetary Bets and Degrees of Belief*, (ii) *Scoring Rules and Degrees of Belief*, (iii) *Axiomatic Approaches to Degrees of Belief*, (iv) *Axiomatic Approaches to Utilities* and (v) *Information Theories*.

For the most part, we shall simply give what seem to us the most important historical references, together with some brief comments. The first two topics will, however, be treated at greater length; partly because of their close relation with the main concerns of this book, and partly because of their connections with the important practical topic of the assessment of beliefs.

Monetary Bets and Degrees of Belief

An elegant demonstration that coherent degrees of belief satisfy the rules of (finitely additive) probability was given by de Finetti (1937/1964), without explicit use of the utility concept. Using the notation for options introduced in Section 2.3, de Finetti's approach can be summarised as follows.

If consequences are assumed to be monetary, and if, given an arbitrary monetary sum m and uncertain event E, an individual's preferences among options are such that $\{pm \mid \Omega\} \sim \{m \mid E, 0 \mid E^c\}$, then the individual's degree of belief in E is defined to be p.

This definition is virtually identical to Bayes' own definition of probability (see our later discussion under the heading of *Axiomatic Approaches to Degrees of Belief*). In modern economic terminology, probability can be considered to be a marginal rate of substitution or, more simply, a kind of "price".

Given that an individual has specified his or her degrees of belief for some collection of events by repeated use of the above definition, either it is possible to arrange a form of monetary bet in terms of these events which is such that the individual will certainly lose, a so-called "Dutch book", or such an arrangement is impossible. In the latter case, the individual is said to have specified a *coherent* set of degrees of belief. It is now straightforward to verify that coherent degrees of belief have the properties of finitely additive probabilities.

To demonstrate that $0 \leq p \leq 1$, for any E and m, we can argue as follows. An individual who assigns $p > 1$ is implicitly agreeing to pay a stake larger than m to enter a gamble in which the maximum prize he or she can win is m; an individual who assigns $p < 0$ is implicitly agreeing to offer a gamble in which he or she will pay out either m or nothing in return for a negative stake, which is equivalent to paying an opponent to enter such a gamble. In either case, a bet can be arranged which will result in a certain loss to the individual and avoidance of this possibility requires that $0 \leq p \leq 1$.

To demonstrate the additive property of degrees of belief for exclusive and exhaustive events, $E_1, E_2, \ldots, E_n$, we proceed as follows. If an individual specifies $p_1, p_2, \ldots, p_n$, to be his or her degrees of belief in those events, this is an implicit agreement to pay a total stake of $p_1 m_1 + p_2 m_2 + \cdots + p_n m_n$ in order to enter a gamble resulting in a prize of m_i if E_i occurs and thus a "gain", or "net return", of $g_i = m_i - \sum_j p_j m_j$, which could, of course, be negative. In order to avoid the possibility of the m_j's being chosen in such a way as to guarantee the negativity of the g_i's for fixed p_j's in this system of linear equations, it is necessary that the determinant of the matrix relating the m_j's to the g_i's be zero so that the linear system cannot be solved; this turns out to require that $p_1 + p_2 + \cdots + p_n = 1$. Moreover, it is easy to check that this is also a sufficient condition for coherence: it implies $\sum_j p_j g_j = 0$, for any choice of the m_j's, and hence the impossibility of all the returns being negative.

The extension of these ideas to cover the revision of degrees of belief conditional on new information proceeds in a similar manner, except that an individual's degree of belief in an event E conditional on an event F is defined to be the number q such that, given any monetary sum m, we have the equivalence $\{qm \,|\, \Omega\} \sim \{m \,|\, E \cap F, 0 \,|\, E^c \cap F, qm \,|\, F^c\}$, according to the individual's preference ordering among options. The interpretation of this definition is straightforward: having paid a stake of qm, if F occurs we are confronted with a gamble with prizes m if E occurs, and nothing otherwise; if F does not occur the bet is "called off" and the stake returned.

However, despite the intuitive appeal of this simple and neat approach, it has two major shortcomings from an operational viewpoint.

In the first place, it is clear that the definitions cannot be taken seriously in terms of arbitrary monetary sums: the "perceived value" of a stake or a return is not equivalent to its monetary value and the missing "utility" concept is required in order to overcome the difficulty. This point was later recognised by de Finetti (see Kyburg and Smokler, 1964/1980, p. 62, footnote (a)), but has its earlier origins in the celebrated *St. Petersburg paradox* (first discussed in terms of utility by Daniel Bernoulli, 1730/1954). For further discussion of possible forms of "utility for money", see, for example, Pratt (1964), Lavalle (1968), Lindley (1971/1985, Chapter 5) and Hull *et al.* (1973). Additionally, one may explicitly recognise that some people have a positive utility for gambling (see, for instance, Conlisk, 1993).

An *ad hoc* modification of de Finetti's approach would be to confine attention to "small" stakes (thus, in effect, restricting attention to a range of outcomes over which the "utility" can be taken as approximately linear) and the argument, thus modified, has considerable pedagogical and, perhaps, practical use, despite its rather informal nature. A more formal argument based on the avoidance of certain losses in betting formulations has been given by Freedman and Purves (1969). Related arguments have also been used by Cornfield (1969), Heath and Sudderth (1972) and Buehler (1976) to expand on de Finetti's concept of coherent systems of bets.

In addition to the problem of "non-linearity in the face of risk", alluded to above, there is also the difficulty that unwanted game-theoretic elements may enter the picture if we base a theory on ideas such as "opponents" choosing the levels of prizes in gambles. For this reason, de Finetti himself later preferred to use an approach based on scoring rules, a concept we have already introduced in Section 2.7.

Scoring Rules and Degrees of Belief

The scoring rule approach to the definition of degrees of belief and the derivation of their properties when constrained to be coherent is due to de Finetti (1963, 1964), with important subsequent generalisations by Savage (1971) and Lindley (1982a).

In terms of the quadratic scoring rule, the development proceeds as follows. Given an uncertain event E, an individual is asked to select a number, p, with the understanding that if E occurs he or she is to suffer a penalty (or loss) of $L = (1 - p)^2$, whereas if E does not occur he or she is to suffer a penalty of $L = p^2$. Using the indicator function for E, the penalty can be written in the general form, $L = (1_E - p)^2$. The number, p, which the individual chooses is defined to be his or her degree of belief in E.

Suppose now that $E_1, E_2, \ldots, E_n$ are an exclusive and exhaustive collection of uncertain events for which the individual, using the quadratic scoring rule scheme, has to specify degrees of belief $p_1, p_2, \ldots, p_n$, respectively, subject now to the penalty

$$L = (1_{E_1} - p_1)^2 + (1_{E_2} - p_2)^2 + \cdots + (1_{E_n} - p_n)^2.$$

Given a specification, $p_1, p_2, \ldots, p_n$, either it is possible to find an alternative specification, $q_1, q_2, \ldots, q_n$, say, such that

$$\sum_{i=1}^{n} (1_{E_i} - q_i)^2 < \sum_{i=1}^{n} (1_{E_i} - p_i)^2,$$

for any assignment of the value 1 to one of the E_i's and 0 to the others, or it is not possible to find such $q_1, q_2, \ldots, q_n$. In the latter case, the individual is said to have specified a *coherent* set of degrees of belief. The underlying idea in this development is clearly very similar to that of de Finetti's (1937/1964) approach where the avoidance of a "Dutch book" is the basic criterion of coherence.

A simple geometric argument now establishes that, for coherence we must have $0 \le p_i \le 1$, for $i = 1, 2, \ldots, n$, and $p_1 + p_2 + \cdots + p_n = 1$. To see this, note that the n logically compatible assignments of values 1 and 0 to the E_i's define n points in $\Re^n$. Thinking of $p_1, p_2, \ldots, p_n$ as defining a further point in $\Re^n$, the coherence condition can be reinterpreted as requiring that this latter point cannot be moved in such a way as to reduce the distance from all the other n points. This means that $p_1, p_2, \ldots, p_n$ must define a point in the convex hull of the other n points, thus establishing the required result.

The extension of this approach to cover the revision of degrees of belief conditional on new information proceeds as follows. An individual's degree of belief in an event E conditional on the occurrence of an event F is defined to be the number q, which he or she chooses when confronted with a penalty defined by $L = 1_F\,(1_E - q)^2$. The interpretation of this penalty is straightforward. Indeed, if F occurs, the specification of q proceeds according to the penalty $(1_E - q)^2$; if F does not occur, there is no penalty, a formulation which is clearly related to the idea of "called-off" bets used in de Finetti's 1937 approach. Suppose now that, in addition to the conditional degree of belief q, the numbers p and r are the individual's degrees of belief, respectively, for the events $E \cap F$ and F, specified subject to the penalty

$$L = 1_F\,(1_E - q)^2 + (1_E\,1_F - p)^2 + (1_F - r)^2.$$

To derive the constraints on p, q and r imposed by coherence, which demands that no other choices will lead to a strictly smaller L, whatever the logically compatible outcomes of the events are, we argue as follows.

If u, v, w, respectively, are the values which L takes in the cases where $E \cap F$, $E^c \cap F$ and F^c occur, then p, q, r satisfy the equations

$$\begin{aligned} u &= (1-q)^2 + (1-p)^2 + (1-r)^2 \\ v &= \quad\; q^2 + \quad\;\; p^2 + (1-r)^2 \\ w &= \qquad\qquad\quad\;\; p^2 + \qquad r^2. \end{aligned}$$

If p, q, r defined a point in $\Re^3$ where the Jacobian of the transformation defined by the above equations did not vanish, it would be possible to move from that point in a direction which simultaneously reduced the values u, v and w. Coherence therefore requires that the Jacobian be zero. A simple calculation shows that this reduces to the condition $q = p/r$, which is, again, Bayes' theorem.

De Finetti's 'penalty criterion' and related ideas have been critically re-examined by a number of authors. Relevant additional references are Myerson (1979), Regazzini (1983), Gatsonis (1984), Eaton (1992) and Gilio (1992a). See, also, Piccinato (1986).

Axiomatic Approaches to Degrees of Belief

Historically, the idea of probability as "degree of belief" has received a great deal of distinguished support, including contributions from James Bernoulli (1713/1899), Laplace (1774/1986, 1814/1952), De Morgan (1847) and Borel (1924/1964). However, so far as we know, none of these writers attempted an axiomatic development of the idea.

The first recognisably "axiomatic" approach to a theory of degrees of belief was that of Bayes (1763) and the magnitude of his achievement has been clearly recognised in the two centuries following his death by the adoption of the adjective *Bayesian* as a description of the philosophical and methodological developments which have been inspired, directly or indirectly, by his essay.

By present day standards, Bayes' formulation is, of course, extremely informal, and a more formal, modern approach only began to emerge a century and a half later, in a series of papers by Wrinch and Jeffreys (1919, 1921). Formal axiom systems which whole-heartedly embrace the principle of revising beliefs through systematic use of Bayes' theorem, are discussed in detail by Jeffreys (1931/1973, 1939/1961), whose profound philosophical and methodological contributions to Bayesian statistics are now widely recognised; see for example, the evaluations of his work by Geisser (1980a), by Good (1980a) and by Lindley (1980a), in the volume edited by Zellner (1980).

From a *foundational perspective*, however, the flavour of Jeffreys' approach seems to us to place insufficient emphasis on the inescapably personal nature of degrees of belief, resulting in an over-concentration on "conventional" representations of degrees of belief derived from "logical" rather than operational considerations (despite the fact that Jeffreys was highly motivated by real world applications!). Similar criticisms seem to us to apply to the original and elegant formal development given by Cox (1946, 1961) and Jaynes (1958), who showed that the probability axioms constitute the only consistent extension of ordinary (Aristotelian) logic in which degrees of belief are represented by real numbers.

We should point out, however, that our emphasis on operational considerations and the subjective character of degrees of belief would, in turn, be criticised by many colleagues who, in other respects, share a basic commitment to the Bayesian approach to statistical problems. See Good (1965, Chapter 2) for a discussion of the variety of attitudes to probability compatible with a systematic use of the Bayesian paradigm.

There are, of course, many other examples of axiomatic approaches to quantifying uncertainty in some form or another. In the finite case, this includes work by Kraft *et al.* (1959), Scott (1964), Fishburn (1970, Chapter 4), Krantz *et al.* (1971), Domotor and Stelzer (1971), Suppes and Zanotti (1976, 1982), Heath and Sudderth (1978) and Luce and Narens (1978). The work of Keynes (1921/1929) and Carnap (1950/1962) deserves particular mention and will be further discussed later in Section 2.8.4. Fishburn (1986) provided an authoritative review of the axiomatic foundations of subjective probability, which is followed by a long, stimulating discussion. See, also, French (1982) and Chuaqui and Malitz (1983).

Axiomatic Approaches to Utilities

Assuming the prior existence of probabilities, von Neumann and Morgenstern (1944/1953) presented axioms for coherent preferences which led to a justification of utilities as numerical measures of value for consequences and to the optimality criterion of maximising expected utility. Much of Savage's (1954/1972) system was directly inspired by this seminal work of von Neumann and Morgenstern and the influence of their ideas extends into a great many of the systems we have mentioned. Other early developments which concentrate on the utility aspects of the decision problem include those of Friedman and Savage (1948, 1952), Marschak (1950), Arrow (1951a), Herstein and Milnor (1953), Edwards (1954) and Debreu (1960). Seminal references are reprinted in Page (1968). General accounts of utility are given in the books by Blackwell and Girshick (1954), Luce and Raiffa (1957), Chernoff and Moses (1959) and Fishburn (1970). Extensive bibliographies are given in Savage (1954/1972) and Fishburn (1968, 1981).

Discussions of the experimental measurement of utility are provided by Edwards (1954), Davison *et al.* (1957), Suppes and Walsh (1959), Becker *et al.* (1963), DeGroot (1963), Becker and McClintock (1967), Savage (1971) and Hull *et al.* (1973). DeGroot (1970, Chapter 7) presents a general axiom system for utilities which imposes rather few mathematical constraints on the underlying decision problem structure. Multiattribute utility theory is discussed, among others, by Fishburn (1964) and Keeney and Raiffa (1976). Other discussions of utility theory include Fishburn (1967a, 1988b) and Machina (1982, 1987). See, also, Schervish *et al.* (1990).

Information Theories

Measures of information are closely related to ideas of uncertainty and probability and there is a considerable literature exploring the connections between these topics.

The logarithmic information measure was proposed independently by Shannon (1948) and Wiener (1948) in the context of communication engineering; Lindley (1956) later suggested its use as a statistical criterion in the design of experiments. The logarithmic divergency measure was first proposed by Kullback and Leibler (1951) and was subsequently used as the basis for an information-theoretic approach to statistics by Kullback (1959/1968). A formal axiomatic approach to measures of information in the context of uncertainty was provided by Good (1966), who has made numerous contributions to the literature of the foundations of decision making and the evaluation of evidence. Other relevant references on information concepts are Renyi (1964, 1966, 1967) and Särndal (1970).

The mathematical results which lead to the characterisation of the logarithmic scoring rule for reporting probability distributions have been available for some considerable time. Logarithmic scores seem to have been first suggested by Good (1952), but he only dealt with dichotomies, for which the uniqueness result is not applicable. The first characterisation of the logarithmic score for a finite distribution was attributed to Gleason by McCarthy (1956); Aczel and Pfanzagl (1966), Arimoto (1970) and Savage (1971) have also given derivations of this form of scoring rule under various regularity conditions.

By considering the inference reporting problem as a particular case of a decision problem, we have provided (in Section 2.7) a natural, unifying account of the fundamental and close relationship between information-theoretic ideas and the Bayesian treatment of “pure inference” problems. Based on work of Bernardo (1979a), this analysis will be extended, in Chapter 3, to cover continuous distributions.

2.8.4 Critical Issues

We shall conclude this chapter by providing a summary overview of our position in relation to some of the objections commonly raised against the foundations of Bayesian statistics. These will be dealt with under the following subheadings: (i) *Dynamic Frame of Discourse*, (ii) *Updating Subjective Probability*, (iii) *Relevance of an Axiomatic Approach*, (iv) *Structure of the Set of Relevant Events*, (v) *Prescriptive Nature of the Axioms*, (vi) *Precise, Complete, Quantitative Preference*, (vii) *Subjectivity of Probability*, (viii) *Statistical Inference as a Decision Problem* and (ix) *Communication and Group Decision Making*.

Dynamic Frame of Discourse

As we indicated in Chapter 1, our concern in this volume is with coherent beliefs and actions in relation to a limited set of specified possibilities, currently assumed necessary and sufficient to reflect key features of interest in the problem under study. In the language of Section 2.2, we are operating in terms of a fixed frame of discourse, defined in the light of our current knowledge and assumptions, M_0. However, as many critics have pointed out, this activity constitutes only one static phase of the wider, evolving, scientific learning and decision process. In the more general, dynamic, context, this activity has to be viewed, either potentially or actually, as sandwiched between two other vital processes. On the one hand, the creative generation of the set of possibilities to be considered; on the other hand, the critical questioning of the adequacy of the currently entertained set of possibilities (see, for example, Box, 1980). We accept that the mode of reasoning encapsulated within the quantitative coherence theory as presented here is ultimately conditional, and thus not directly applicable to every phase of the scientific process. But we do not accept, as Box (1980) appears to, that alternative *formal* statistical theories have a convincing, complementary role to play.

The problem of generating the frame of discourse, i.e., inventing new models or theories, seems to us to be one which currently lies outside the purview of any “statistical” formalism, although some limited formal clarification is actually possible within the Bayesian framework, as we shall see in Chapter 4. Substantive subject-matter inputs would seem to be of primary importance, although informal, exploratory data analysis is no doubt a necessary adjunct and, particularly in the context of the possibilities opened up by modern computer graphics, offers considerable intellectual excitement and satisfaction in its own right.

The problem of criticising the frame of discourse also seems to us to remain essentially unsolved by any "statistical" theory. In the case of a "revolution", or even "rebellion", in scientific paradigm (Kuhn, 1962), the issue is resolved for us as statisticians by the consensus of the subject-matter experts, and we simply begin again on the basis of the frame of discourse implicit in the new paradigm. However, in the absence of such "externally" directed revision or extension of the current frame of discourse, it is not clear what questions one should pose in order to arrive at an "internal" assessment of adequacy in the light of the information thus far available.

On the one hand, exploratory diagnostic probing would seem to have a role to play in confirming that specific forms of local elaboration of the frame of discourse should be made. The logical catch here, however, is that such specific diagnostic probing can only stem from the prior realisation that the corresponding specific elaborations might be required. The latter could therefore be incorporated *ab initio* into the frame of discourse and a fully coherent analysis carried out. The issue here is one of pragmatic convenience, rather than of circumscribing the scope of the coherent theory.

On the other hand, the issue of assessing adequacy in relation to a *total absence of any specific suggested elaborations* seems to us to remain an open problem. Indeed, it is not clear that the "problem" as usually posed is well-formulated. For example, is the key issue that of "surprise"; or is some kind of extension of the notion of a decision problem required in order to give an operational meaning to the concept of "adequacy"?

Readers interested in this topic will find in Box (1980), and the ensuing discussion, a range of reactions. We shall return to these issues in Chapter 6. Related issues arise in discussions of the general problem of assessing, or "calibrating", the external, empirical performance of an internally coherent individual; see, for example, Dawid (1982a).

Overall, our responses to critics who question the relevance of the coherent approach based on a fixed frame of reference can be summarised as follows. So far as the scope and limits of Bayesian theory are concerned: (i) we acknowledge that the mode of reasoning encapsulated within the quantitative coherence theory is ultimately *conditional*, and thus not directly applicable to every phase of the scientific process; (ii) informal, *exploratory* techniques are an essential part of the process of generating ideas; there can be no purely "statistical" theory of model formulation; this aspect of the scientific process is not part of the foundational debate, although the process of passing from such ideas to their mathematical *representation* can often be subjected to formal analysis; (iii) we *all* lack a decent theoretical formulation of and solution to the problem of global model criticism in the absence of concrete suggested alternatives.

However, critics of the Bayesian approach should recognise that: (i) an enormous amount of current theoretical and applied statistical activity is concerned with the analysis of uncertainty in the context of models which are accepted, for the purposes of the analysis, as working frames of discourse, subject only to local probing of specific potential elaborations, and (ii) our arguments thus far, and those to follow, are an attempt to convince the reader that *within this latter context* there are compelling reasons for adopting the Bayesian approach to statistical theory and practice.

Updating Subjective Probability

An issue related to the topic just discussed is that of the mechanism for updating subjective probabilities.

In Section 2.4.2, we defined, in terms of a conditional uncertainty relation, $\leq_G$, the notion of the conditional probability, $P(E \mid G)$, of an event E given the *assumed* occurrence of an event G. From this, we derived Bayes' theorem, which establishes that $p(E \mid G) = P(G \mid E)P(E)/P(G)$. If we actually *know for certain* that G has occurred, $P(E \mid G)$ becomes our actual degree of belief in E. The *prior* probability $P(E)$, has been updated to the *posterior* probability $P(E \mid G)$.

However, a number of authors have questioned whether it is justified to identify assessments made conditional on the *assumed* occurrence of G with actual beliefs once G is *known*. We shall not pursue this issue further, although we acknowledge its interest and potential importance. Detailed discussion and relevant references can be found in Diaconis and Zabell (1982), who discuss, in particular, *Jeffrey's rule* (Jeffrey, 1965/1983), and Goldstein (1985), who examines the role of *temporal coherence*. See, also, Good (1977).

Relevance of the Axiomatic Approach

Arguments against over-concern with foundational issues come in many forms. At one extreme, we have heard Bayesian colleagues argue that the mechanics and flavour of the Bayesian inference process have their own sufficient, direct, intuitive appeal and do not need axiomatic reinforcement. Another form of this argument asserts that developments from axiom systems are "pointless" because the conclusions are, tautologically, contained in the premises. Although this is literally true, we simply do not accept that the methodological imperatives which flow from the assumptions of quantitative coherence are in any way "obvious" to someone contemplating the axioms. At the other extreme, we have heard proponents of supposedly "model-free" exploratory methodology proclaim that we can evolve towards "good practice" by simply giving full encouragement to the creative imagination and then "seeing what works".

Our objection to both these attitudes is that they each implicitly assume, albeit from different perspectives, the existence of a commonly agreed notion of what constitutes "desirable statistical practice". This does not seem to us a reasonable assumption at all, and to avoid potential confusion, an operational definition of the notion is required. The quantitative coherence approach is based on the assumption that, *within the structured framework set out in Section 2.2*, desirable practice requires, at least, to avoid Dutch-book inconsistencies, an assumption which leads to the Bayesian paradigm for the revision of belief.

Structure of the Set of Relevant Events

But is the structure assumed for the set of relevant events too rigid? In particular, is it reasonable to assume that, in each and every context involving uncertainty, the logical description of the possibilities should be forced into the structure of an algebra (or σ-algebra), in which each event has the same logical status? It seems to us that this may not always be reasonable and that there is a potential need for further research into the implications of applying appropriate concepts of quantitative coherence to event structures other than simple algebras. For example, this problem has already been considered in relation to the foundations of quantum mechanics, where the notion of "sample space" has been generalised to allow for the simultaneous representation of the outcomes of a set of "related" experiments (see, for example, Randall and Foulis, 1975). In that context, it has been established that there exists a natural extension of the Bayesian paradigm to the more general setting.

Another area where the applicability of the standard paradigm has been questioned is that of so-called "knowledge-based expert systems", which often operate on knowledge representations which involve complex and loosely structured spaces of possibilities, including hierarchies and networks. Proponents of such systems have argued that (Bayesian) probabilistic reasoning is incapable of analysing these structures and that novel forms of quantitative representations of uncertainty are required (see Spiegelhalter and Knill-Jones, 1984, and ensuing discussion, for references to these ideas). However, alternative proposals, which include "fuzzy logic", "belief functions" and "confirmation theory", are, for the most part, *ad hoc* and the challenge to the probabilistic paradigm seems to us to be elegantly answered by Lauritzen and Spiegelhalter (1988). We shall return to this topic later in this section.

Finally, another form of query relating to the logical status of events is sometimes raised (see, for example, Barnard, 1980a). This draws attention to the interpretational asymmetry between a statement like "the underlying distribution is normal" and its negation. This raises questions about their implicitly symmetric treatment within the framework given in Section 2.2. Choices of the elements to be included in $\mathcal{E}$ are, of course, bound up with general questions of "modelling" and the issue here seems to us to be one concerning sensible modelling strategies. We shall return to this topic in Chapters 4 and 6.

Prescriptive Nature of the Axioms

When introducing our formal development, we emphasised that the Bayesian foundational approach is prescriptive and not descriptive. We are concerned with understanding how we *ought* to proceed, *if* we wish to avoid a specified form of behavioural inconsistency. We are *not* concerned with sociological or psychological description of actual behaviour. For the latter, see, for example, Wallsten (1974), Kahneman and Tversky (1979), Kahneman *et al.* (1982), Machina (1987), Bordley (1992), Luce (1992) and Yilmaz (1992). See, also, Savage (1980).

Despite this, many critics of the Bayesian approach have somehow taken comfort from the fact that there is empirical evidence, from experiments involving hypothetical gambles, which suggests that people often do not act in conformity with the coherence axioms; see, for example, Allais (1953) and Ellsberg (1961).

Allais' criticism is based on a study of the actual preferences of individuals in contexts where they are faced with pairs of hypothetical situations, like those described in Figure 2.8, in each of which a choice has to be made between the two options where C stands for current assets and the numbers describe thousands of units of a familiar currency.

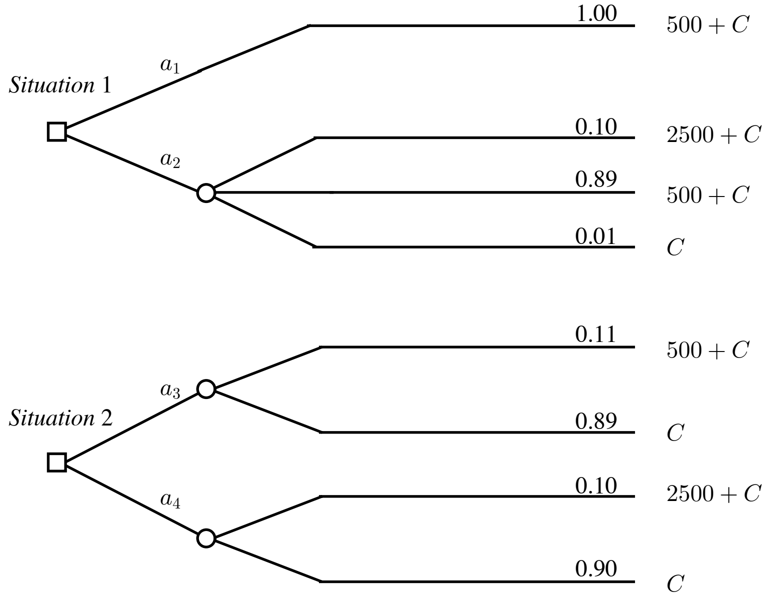

Figure 2.8 *An illustration of Allais' paradox*

It has been found (see, for example, Allais and Hagen, 1979) that there are a great many individuals who prefer option 1 to option 2 in the first situation, and at the same time prefer option 4 to option 3 in the second situation.

To examine the coherence of these two revealed preferences, we note that, if they are to correspond to a consistent utility ordering, there must exist a utility function $u(.)$, defined over consequences (in this case, total assets in thousands of monetary units), satisfying the inequalities

$$u(500 + C) > 0.10\,u(2,500 + C) + 0.89\,u(500 + C) + 0.01\,u(C)$$

$$0.10\,u(2,500 + C) + 0.90\,u(C) > 0.11\,u(500 + C) + 0.89\,u(C).$$

But simple rearrangement reveals that these inequalities are logically incompatible for any function $u(.)$, and, therefore, the stated preferences are incoherent.

How should one react to this conflict between the compelling intuitive attraction (for many individuals) of the originally stated preferences, and the realisation that they are not in accord with the prescriptive requirements of the formal theory? Allais and his followers would argue that the force of examples of this kind is so powerful that it undermines the whole basis of the axiomatic approach set out in Section 2.3. This seems to us a very peculiar argument. It is as if one were to argue for the abandonment of ordinary logical or arithmetic rules, on the grounds that individuals can often be shown to perform badly at deduction or long division.

The conclusion to be drawn is surely the opposite: namely, the more liable people are to make mistakes, the more need there is to have the formal prescription available, both as a reference point, to enable us to discover the kinds of mistakes and distortions to which we are prone in *ad hoc* reasoning, and also as a suggestive source of improved strategies for thinking about and structuring problems.

Table 2.4 *Savage's reformulation of Allais' example*

	Ticket number	1	2–11	12–100
situation 1	*option* 1	$500 + C$	$500 + C$	$500 + C$
	option 2	C	$2500 + C$	$500 + C$
situation 2	*option* 3	$500 + C$	$500 + C$	C
	option 4	C	$2500 + C$	C

In the case of Allais' example, Savage (1954/1972, Chapter 5) pointed out that a concrete realisation of the options described in the two situations could be achieved by viewing the outcomes as prizes from a lottery involving one hundred numbered tickets, as shown in Table 2.4. Indeed, when the problem is set out in this form, it is clear that if any of the tickets numbered from 12 to 100 is chosen it will not matter, in either situation, which of the options is selected. Preferences in both situations should therefore only depend on considerations relating to tickets in the range from 1 to 11. But, for this range of tickets, situations 1 and 2 are identical in structure, so that preferring option 1 to option 2 and at the same time preferring option 4 to option 3 is now seen to be indefensible.

Viewed in this way, Allais' problem takes on the appearance of a decision-theoretic version of an "optical illusion" achieved through the distorting effects of "extreme" consequences, which go far beyond the ranges of our normal experience. The lesson of Savage's analysis is that, when confronted with complex or tricky problems, we must be prepared to shift our angle of vision in order to view the structure in terms of more concrete and familiar images with which we feel more comfortable.

Ellsberg's (1961) criticism is of a similar kind to Allais', but the "distorting" elements which are present in his hypothetical gambles stem from the rather vague nature of the uncertainty mechanisms involved, rather than from the extreme nature of the consequences. In such cases, where confusion is engendered by the probabilities rather than the utilities, the perceived incoherence may, in fact, disappear if one takes into account the possibility that the experimental subjects' utility may be a function of more than one attribute. In particular, we may need to consider the attribute "avoidance of looking foolish", often as a result of thinking that there is a "right answer" if the problem seems predominantly to do with sorting out "experimentally assigned" probabilities, in addition to the monetary consequences specified in the hypothetical gambles. Even without such refinements, however, and arguing solely in terms of the gambles themselves, Raiffa (1961) and Roberts (1963) have provided clear and convincing rejoinders to the Ellsberg criticism. Indeed, Roberts presents a particularly lucid and powerful defence of the axioms, also making use of the analogy with "optical" and "magical" illusions. The form of argument used is similar to that in Savage's rejoinder to Allais, and we shall not repeat the details here. For a recent discussion of both the Allais and Ellsberg phenomena, see Kadane (1992).

Precise, Complete, Quantitative Preferences

In our axiomatic development we have not made the a priori assumption that all options can be compared directly using the preference relation. We have, however, assumed, in Axiom 5, that all consequences and certain general forms of dichotomised options can be compared with dichotomised options involving standard events. This latter assumption then turns out to imply a quantitative basis for all preferences, and hence for beliefs and values.

The view has been put forward by some writers (e.g. Keynes, 1921/1929, and Koopman, 1940) that not all degrees of belief are quantifiable, or even comparable. However, beginning with Jeffreys' review of Keynes' *Treatise* (see also Jeffreys, 1931/1973) the general response to this view has been that some form of quantification is essential if we are to have an operational, scientifically useful theory. Other references, together with a thorough review of the mathematical consequences of these kind of assumptions, are given by Fine (1973, Chapter 2).

Nevertheless, there has been a widespread feeling that the demand for precise quantification, implicit in "standard" axiom systems, is rather severe and certainly ought to be questioned. We should consider, therefore, some of the kinds of suggestions that have been put forward from this latter perspective.

Among the attempts to present formal alternatives to the assumption of precise quantification are those of Good (1950, 1962), Kyburg (1961), Smith (1961), Dempster (1967, 1985), Walley and Fine (1979), Girón and Ríos (1980), DeRobertis and Hartigan (1981), Walley (1987, 1991) and Nakamura (1993). In essence, the suggestion in relation to probabilities is to replace the usual representation of a degree of belief in terms of a single number, by an interval defined by two numbers, to be interpreted as "upper" and "lower" probabilities. So far as decisions are concerned, such theories lead to the identification of a class of "would-be" actions, but provide no operational guidance as to how to choose from among these. Particular ideas, such as Dempster's (1968) generalization of the Bayesian inference mechanism, have been shown to be suspect (see, for example, Aitchison, 1968), but have led on themselves to further generalizations, such as Shafer's (1976, 1982a) theory of "belief functions". This has attracted some interest (see e.g., Wasserman (1990a, 1990b), but its operational content has thus far eluded us.

In general, we accept that the assumption of precise quantification, i.e., that comparisons with standard options can be successively refined without limit, is clearly absurd if taken literally and interpreted in a *descriptive* sense. We therefore echo our earlier detailed commentary on Axiom 5 in Section 2.3, to the effect that these kinds of proposed extension of the axioms seem to us to be based on a confusion of the descriptive and the prescriptive and to be largely unnecessary. It is rather as though physicists and surveyors were to feel the need to rethink their practices on the basis of a physical theory incorporating explicit concepts of upper and lower lengths. We would not wish, however, to be dogmatic about this. Our basic commitment is to quantitative coherence. The question of whether this should be precise, or allowed to be imprecise, is certainly an open, debatable one, and it might well be argued that "measurement" of beliefs and values is not totally analogous to that of physical "length". An obvious, if often technically involved solution, is to consider simultaneously all probabilities which are compatible with elicited comparisons. This and other forms of "robust Bayesian" approaches will be reviewed in Section 5.6.3. In this work, we shall proceed on the basis of a *prescriptive* theory which assumes precise quantification, but then pragmatically acknowledges that, in practice, all this should be taken with a large pinch of salt and a great deal of systematic sensitivity analysis. For a related practical discussion, see Hacking (1965). See, also, Chateaneuf and Jaffray (1984).

Subjectivity of Probability

As we stressed in Section 2.2, the notion of preference between options, the primitive operational concept which underlies all our other definitions, is to be understood as personal, in the sense that it derives from the response of a particular individual to a decision making situation under uncertainty. A particular consequence of this is that the concept which emerges is personal degree of belief, defined in Section 2.4 and subsequently shown to combine for compound events in conformity with the properties of a finitely additive probability measure.

The "individual" referred to above could, of course, be some kind of group, such as a committee, provided the latter had agreed to "speak with a single voice", in which case, to the extent that we ignore the processes by which the group arrives at preferences, it can conveniently be regarded as a "person". Further comments on the problem of individuals versus groups will be given later under the heading *Communication and Group Decision Making*.

This idea that personal (or subjective) probability should be the key to the "scientific" or "rational" treatment of uncertainty has proved decidedly unpalatable to many statisticians and philosophers (although in some application areas, such as actuarial science, it has met with a more favourable reception; see Clarke, 1954). At the very least, it appears to offend directly against the general notion that the methods of science should, above all else, have an "objective" character. Nevertheless, bitter though the subjectivist pill may be, and admittedly difficult to swallow, the alternatives are either inert, or have unpleasant and unexpected side-effects or, to the extent that they appear successful, are found to contain subjectivist ingredients.

From the objectivistic standpoint, there have emerged two alternative kinds of approach to the definition of "probability" both seeking to avoid the subjective degree of belief interpretation. The first of these retains the idea of probability as measurement of partial belief, but rejects the subjectivist interpretation of the latter, regarding it, instead, as a unique degree of partial *logical* implication between one statement and another. The second approach, by far the most widely accepted in some form or another, asserts that the notion of probability should be related in a fundamental way to certain "objective" aspects of physical reality, such as *symmetries or frequencies*.

The logical view was given its first explicit formulation by Keynes (1921/1929) and was later championed by Carnap (1950/1962) and others; it is interesting to note, however, that Keynes seems subsequently to have changed his view and acknowledged the primacy of the subjectivist interpretation (see Good, 1965, Chapter 2). Brown (1993) proposes the related concept of "impersonal" probability.

From an historical point of view, the first systematic foundation of the frequentist approach is usually attributed to Venn (1886), with later influential contributions from von Mises (1928) and Reichenbach (1935). The case for the subjectivist approach and against the objectivist alternatives can be summarised as follows.

The *logical* view is entirely lacking in operational content. Unique probability values are simply assumed to exist as a measure of the degree of implication between one statement and another, to be intuited, in some undefined way, from the formal structure of the language in which these statements are presented.

The *symmetry* (or classical) view asserts that physical considerations of symmetry lead directly to a primitive notion of "equally likely cases". But any uncertain situation typically possesses many plausible "symmetries": a truly "objective" theory would therefore require a procedure for choosing a particular symmetry and for justifying that choice. The subjectivist view explicitly recognises that regarding a specific symmetry as probabilistically significant is itself, inescapably, an act of *personal* judgement.

The *frequency* view can only attempt to assign a measure of uncertainty to an individual event by embedding it in an infinite class of "similar" events having certain "randomness" properties, a "collective" in von Mises' (1928) terminology, and then identifying "probability" with some notion of limiting relative frequency. But an individual event can be embedded in many different "collectives" with no guarantee of the same resulting limiting relative frequencies: a truly "objective" theory would therefore require a procedure for justifying the choice of a particular embedding sequence. Moreover, there are obvious difficulties in defining the underlying notions of "similar" and "randomness" without lapsing into some kind of circularity. The subjectivist view explicitly recognises that any assertion of "similarity" among different, individual events is itself, inescapably, an act of *personal* judgement, requiring, in addition, an operational definition of which is meant by "similar".

In fact, this latter requirement finds natural expression in the concept of an *exchangeable* sequence of events, which we shall discuss at length in Chapter 4. This concept, via the celebrated de Finetti representation theorem, provides an elegant and illuminating explanation, from an entirely subjectivistic perspective, of the fundamental role of symmetries and frequencies in the structuring and evaluation of personal beliefs. It also provides a meaningful operational interpretation of the word "objective" in terms of "intersubjective consensus".

The identification of probability with frequency or symmetry seems to us to be profoundly misguided. It is of paramount importance to maintain the distinction between the *definition of a general concept* and the *evaluation of a particular case*. In the subjectivist approach, the definition derives from logical notions of quantitative coherent preferences: practical evaluations in particular instances often derive from perceived symmetries and observed frequencies, and it is only in this evaluatory process that the latter have a role to play.

The subjectivist point of view outlined above is, course, not new and has been expounded at considerable length and over many years by a number of authors. The idea of probability as individual "degree of confidence" in an event whose outcome is uncertain seems to have been first put forward by James Bernoulli (1713/1899). However, it was not until Thomas Bayes' (1763) famous essay that it was explicitly used as a definition:

> The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.

Not only is this directly expressed in terms of operational comparisons of certain kinds of simple options on the basis of expected values, but the style of Bayes' presentation strongly suggests that these expectations were to be interpreted as personal evaluations.

A number of later contributions to the field of subjective probability are collected together and discussed in the volume edited by Kyburg and Smokler (1964/1980), which includes important seminal papers by Ramsey (1926) and de Finetti (1937/1964). An exhaustive and profound discussion of all aspects of subjective probability is given in de Finetti's magisterial *Theory of Probability* (1970/1974, 1970/1975). Other interpretations of probability are discussed in Renyi (1955), Good (1959), Kyburg (1961, 1974), Fishburn (1964), Fine (1973), Hacking (1975), de Finetti (1978), Walley and Fine (1979) and Shafer (1990).

Statistical Inference as a Decision Problem

Stylised statistical problems have often been approached from a decision-theoretical viewpoint; see, for instance, the books by Ferguson (1967), DeGroot (1970), Barnett (1973/1982), Berger (1985a) and references therein. However, we have already made clear that, in our view, the supposed dichotomy between inference and decision is illusory, since any report or communication of beliefs following the receipt of information inevitably itself constitutes a form of action. In Section 2.7, we formalised this argument and characterised the utility structure that is typically appropriate for consequences in the special case of a "pure inference" problem. The expected utility of an "experiment" in this context was then seen to be identified with expected information (in the Shannon sense), and a number of information-theoretic ideas and their applications were given a unified interpretation within a purely subjectivist Bayesian framework.

Many approaches to statistical inference do not, of course, assign a primary role to reporting probability distributions, and concentrate instead on stylised estimation and hypothesis testing formulations of the problem (see Appendix B, Section 3). We shall deal with these topics in more detail in Chapters 5 and 6.

Communication and Group Decision Making

The Bayesian approach which has been presented in this chapter is predicated on the primitive notion of *individual* preference. A seemingly powerful argument against the use of the Bayesian paradigm is therefore that it provides an inappropriate basis for the kinds of interpersonal communication and reporting processes which characterise both public debate about beliefs regarding scientific and social issues, and also "cohesive-small-group" decision making processes. We believe that the two contexts, "public" and "cohesive-small-group", pose rather different problems, requiring separate discussion.

In the case of the revision and communication of beliefs in the context of general scientific and social debate, we feel that criticism of the Bayesian paradigm is largely based on a misunderstanding of the issues involved, and on an over-simplified view of the paradigm itself, and the uses to which it can be put. So far as the issues are concerned, we need to distinguish two rather different activities: on the one hand, the prescriptive processes by which we ought individually to revise our beliefs in the light of new information if we aspire to coherence; on the other hand, the pragmatic processes by which we seek to report to and share perceptions with others. The first of these processes leads us inescapably to the conclusion that beliefs should be handled using the Bayesian paradigm; the second reminds us that a "one-off" application of the paradigm to summarise a single individual's revision of beliefs is inappropriate in this context.

But, so far as we are aware, no Bayesian statistician has ever argued that the latter would be appropriate. Indeed, the whole basis of the subjectivist philosophy predisposes Bayesians to seek to report a rich *range* of the possible belief mappings induced by a data set, the range being chosen both to reflect (and even to challenge) the initial beliefs of a range of interested parties. Some discussion of the Bayesian reporting process may be found in Dickey (1973), Dickey and Freeman (1975) and Smith (1978). Further discussion is given in Smith (1984), together with a review of the connections between this issue and the role of models in facilitating communication and consensus. This latter topic will be further considered in Chapter 4.

We concede that much remains to be done in developing Bayesian reporting technology, and we conjecture that modern interactive computing and graphics will have a major role to play. Some of the literature on expert systems is relevant here; see, for instance, Lindley (1987), Spiegelhalter (1987) and Gaul and Schader (1988). On the broader issue, however, one of the most attractive features of the Bayesian approach is its recognition of the legitimacy of the plurality of (coherently constrained) responses to data. Any approach to scientific inference which seeks to legitimise *an answer* in response to complex uncertainty seems to us a totalitarian parody of a would-be rational human learning process.

On the other hand, in the "cohesive-small-group" context there may be an imposed need for *group* belief and decision. A variety of problems can be isolated within this framework, depending on whether the emphasis is on combining probabilities, or utilities, or both; and on how the group is structured in relation to such issues as "democracy", "information-sharing", "negotiation" or "competition". It is not yet clear to us whether the analyses of these issues will impinge directly on the broader controversies regarding scientific inference methodology, and so we shall not attempt a detailed review of the considerable literature that is emerging.

Useful introductions to the extensive literature on amalgamation of beliefs or utilities, together with most of the key references, are provided by Arrow (1951b), Edwards (1954), Luce and Raiffa (1957), Luce (1959), Stone (1961), Blackwell and Dubins (1962), Fishburn (1964, 1968, 1970, 1987), Kogan and Wallace (1964), Wilson (1968), Winkler (1968, 1981), Sen (1970), Kranz *et al.* (1971), Marschak and Radner (1972), Cochrane and Zeleny (1973), DeGroot (1974, 1980), Morris (1974), White and Bowen (1975), White (1976a, 1976b), Press (1978, 1980b, 1985b), Lindley *et al.* (1979), Roberts (1979), Hogarth (1980), Saaty (1980), Berger (1981), French (1981, 1985, 1986, 1989), Hylland and Zeckhauser (1981), Weerahandi and Zidek (1981, 1983), Brown and Lindley (1982, 1986), Chankong and Haimes (1982), Edwards and Newman (1982), DeGroot and Feinberg (1982, 1983, 1986), Raiffa (1982), French *et al.* (1983), Lindley (1983, 1985, 1986), Bunn (1984), Caro *et al.* (1984), Genest (1984a, 1984b), Yu (1985), De Waal *et al.* (1986), Genest and Zidek (1986), Arrow and Raynaud (1987), Clemen and Winkler (1987, 1993), Kim and Roush (1987), Barlow *et al.* (1988), Bayarri and DeGroot (1988, 1989, 1991), Huseby (1988), West (1988, 1992a), Clemen (1989, 1990), Ríos *et al.* (1989), Seidenfeld *et al.* (1989), Ríos (1990), DeGroot and Mortera (1991), Kelly (1991), Lindley and Singpurwalla (1991, 1993), Goel *et al.* (1992), Goicoechea *et al.* (1992), Normand and Tritchler (1992) and Gilardoni and Clayton (1993). Important, seminal papers are reproduced in Gärdenfors and Sahlin (1968). For related discussion in the context of policy analysis, see Hodges (1987).

References relating to the Bayesian approach to game theory include Harsany (1967), DeGroot and Kadane (1980), Eliashberg and Winkler (1981), Kadane and Larkey (1982, 1983), Raiffa (1982), Wilson (1986), Aumann (1987), Smith (1988b), Nau and McCardle (1990), Young and Smith (1991), Kadane and Seidenfeld (1992) and Keeney (1992).

A recent review of related topics, followed by an informative discussion, is provided by Kadane (1993).