Chapter 3

Quantifying Uncertainty

When it comes to reasoning about likelihood, we often do not want to say only that one set or event is more likely than another; we want to quantify how much more likely. This typically means assigning a number to the likelihood. But then, of course, we must explain what the numbers mean.

For those steeped in probability, there is only one appropriate model for numeric uncertainty, and that is probability. However, there are other approaches that are worth considering. In this chapter, I consider five more quantitative approaches to representing uncertainty. The first is probability. The second uses what are called *Dempster-Shafer belief functions*. The third uses *possibility measures*, which are based on fuzzy logic. The fourth is a way of measuring "degree of surprise" in terms of what are called *ranking functions*. The first four notions all attempt to quantify uncertainty using numbers, either real numbers (in the first three cases) or integers (in the case of ranking functions). One disadvantage of using numbers is that we are forced to totally order likelihood: For any two sets A and B, either A is at least as likely than B or B is at least as likely as A. The fifth notion uses *plausibility measures*; it does not use numbers and thus allows uncertainty to be partially ordered. As we shall see, it can be viewed as generalizing all the other notions we consider.

Each of the approaches I consider in this chapter has associated with it a notion of *updating*, which describes how a measure should be updated in the light of additional information. The focus in this chapter is on issues of interpretation; in the next chapter I consider how beliefs should be updated.

3.1 Probability Measures

Most readers have probably seen probability before, so I do not go into great detail here. However, I do try to give enough of a review of probability to make the presentation completely self-contained. Even readers familiar with this material may want to scan it briefly, just to get used to the notation.

Suppose first that we have a finite set W of possible worlds, say $W = \{w_1, \ldots, w_n\}$. We can then assign to each of these worlds a number—a *probability*—that we can think of as describing the likelihood of that world being the actual world. For example, if we are tossing a die, then there are six possible worlds (sometimes called *elementary outcomes*); we can think of world w_i as being the world where the die landed i, for $i = 1, \ldots, 6$. We might assign each of these worlds the number 1/6, if we think each of the outcomes is equally likely. (The choice of 1/6 is made so that the sum is 1; see below.) On the other hand, if the outcome of 1 is much more likely than the others, we might assign w_1 probability 1/2, and all the other outcomes probability 1/10.

Assuming that we attach probability 1/6 to each elementary outcome, what probability should we attach to the die landing either 1 or 2? This can be thought of as the probability of the set $\{w_1, w_2\}$. It seems reasonable to take the answer to be 1/3, the sum of the probability of landing 1 and the probability of landing 2. Thus, the probability of the whole space $\{w_1, \ldots, w_6\}$, is 1, the sum of the probabilities of all the possible outcomes. In probability theory, 1 is conventionally taken to denote certainty. Since it is certain that there will be some outcome, we want the probability of the whole space to be 1. That is why we chose 1/6 for the probability of each of the six elementary outcomes.

This discussion shows that, even though we may start by thinking of probabilities of worlds, ultimately, we want to assign probabilities to *sets* (i.e. *events*). (Recall that this issue also arose in the context of relative likelihood in Section 2.3.) In most of the examples in this book all the subsets of a set of worlds are assigned a probability. Nevertheless, there are good reasons, both technical and philosophical, not to *require* that a probability measure be defined on all sets. There are some constraints on the sets on which a probability measure is defined. If we can talk about the probability of A and B, we also want to be able to talk about the probability of $A \cup B$ and the probability of $\overline{A}$.

Definition 3.1.1 An *algebra* of subsets of W is a set of subsets of W that contains W and is closed under union and complementation. ∎

Thus, a set $\mathcal{F}$ of subsets of W is an algebra if, whenever U and V are in

$\mathcal{F}$, then so are $U \cup V$ and $\overline{U}$. (Note that this means that an algebra is also closed under intersection as well.) The domain of a probability measure is an algebra of sets. (In the case that W is infinite, it is often required to be a *σ-algebra*, that is, closed under countable unions as well as finite unions.) By convention, the range of a probability measure is the interval $[0, 1]$. (In general, $[a, b]$ denotes the set of reals between a and b, including both a and b, that is, $[a, b] = \{x \in I\!R : a \le x \le b\}$.)

Definition 3.1.2 A *probability space* is a tuple $(W, \mathcal{F}, \mu)$, where $\mathcal{F}$ is a an algebra of subsets of W and $\mu : \mathcal{F} \to [0, 1]$ satisfies the following two properties:

P1. $\mu(W) = 1$.

P2. $\mu(U \cup V) = \mu(U) + \mu(V)$ if U and V are disjoint elements of $\mathcal{F}$. ∎

The sets in $\mathcal{F}$ are called the *measurable sets*; μ is called a *probability measure on* W (or on $\mathcal{F}$, if we want to emphasize the domain of μ). W is often called a *sample space* in probability texts. Notice that the arguments to μ are not elements of W but subsets of W. If the argument is a singleton subset $\{w\}$, I often abuse notation and write $\mu(w)$ rather than $\mu(\{w\})$. This convention is also followed for the other notions of uncertainty introduced in this chapter.

It follows from P1 and P2 that $\mu(\emptyset) = 0$: Since $\emptyset$ and W are disjoint,

$$1 = \mu(W) = \mu(W \cup \emptyset) = \mu(W) + \mu(\emptyset) = 1 + \mu(\emptyset),$$

so $\mu(\emptyset) = 0$.

Although P2 applies only to pairs of sets, an easy induction argument also shows that if $U_1, \ldots, U_k$ are pairwise disjoint elements of $\mathcal{F}$, then

$$\mu(U_1 \cup \ldots \cup U_k) = \mu(U_1) + \cdots + \mu(U_k).$$

This property is known as *finite additivity*. Note that it follows that if W is finite and $\mathcal{F}$ consists of all subsets of W, then a probability measure can be characterized as a function $\mu : W \to [0, 1]$ such that $\sum_{w \in W} \mu(w) = 1$. That is, we can take the domain of the probability measure to be the elements in W, and then extend μ to subsets by taking $\mu(U) = \sum_{u \in U} \mu(u)$. While the assumption that all sets are measurable is certainly an important special case, I have taken the more traditional definition, with the domain of μ just the sets in $\mathcal{F}$, since it allows greater generality.

If W is infinite and $\mathcal{F}$ is *σ-algebra* (so that it is closed under countable unions as well as finite unions), it is typically required that μ is *σ-additive* or *countably additive*, so that if $U_1, U_2, \ldots$ are disjoint elements of $\mathcal{F}$, then $\mu(\cup_i U_i) = \mu(U_1) + \mu(U_2) + \cdots$. Since, in most cases, we shall be interested in situations where the set of possible worlds is finite, I largely ignore the issue of whether probability is countably additive or only finitely additive.

3.1.1 Justifying Probability

If we quantify our beliefs using probability, we must be careful to explain what the numbers represent, where they come from, and why finite additivity is appropriate. Without such an explanation, it will not be clear how to assign probabilities in applications, nor how to interpret the results we get when using probability.

The classical approach to applying probability, which goes back to the 17th and 18th centuries, is to reduce the situation we want to analyze to a number of "basic" outcomes. We then apply the *principle of indifference* and take all these outcomes to be equally likely. Intuitively, the basic outcomes ought to be chosen so that, in the absence of any other information, we have no reason to consider one more likely than another. In terms of our framework, the possible worlds correspond to the possible outcomes. If there are n basic outcomes, the probability of each one is $1/n$. The probability of a set of k outcomes is k/n. Clearly this definition satisfies P1 and P2.

This is certainly the justification for ascribing to each of the six outcomes of the toss of a die a probability of 1/6. By using powerful techniques of combinatorics together with the principle of indifference, card players can compute the probability of getting various kinds of hands, and then use this information to guide their play of the game. Moreover, the principle of indifference essentially describes how we handle situations with statistical information. For example, if we know that 40% of a doctor's patients are over 60, and a nurse informs the doctor that one of his patients is waiting for him in the waiting room, it seems reasonable for the doctor to say that the likelihood of that patient being over 60 is .4. Essentially what is going on here is that there is one possible world for each of the possible patients that might be in the waiting room. If we take all these worlds to be equally probable, then the probability of the patient being over 60 will indeed be .4. (I return to the principle of indifference and the relationship between statistical information and probability in Chapter **??**.)

While taking possible worlds to be equally probable is a very compelling intuition, the trouble with the principle of indifference is that it is not always obvious how to reduce a situation into basic outcomes that seem equally likely. This is a significant concern, because different choices of basic outcomes will in general lead to different answers. For example, if we are asked the probability that a couple with two children has two boys, the most obvious way of applying the principle of indifference would suggest that the answer is 1/3. After all, the two children could be either (a) two boys, (b) two girls, or (c) a boy and a girl. If all these outcomes are equally likely, then the probability of having two boys is 1/3.

There is, however, another way of applying the principle of indifference. We can describe the possible outcomes as (B, B), (B, G), (G, B), and (G, G): (a) both children are boys, (b) the first child is a boy and the second a girl, (c) the first child is a girl and the second a boy, and (d) both children are girls. If we apply the principle of indifference to this description of the possible outcomes, the probability of having two boys is 1/4.

The latter answer accords better with observed frequencies, and there are compelling general reasons to consider the second approach better than the first for constructing the set of possible outcomes. But in many other cases, it is far from obvious how to choose the basic outcomes. And after we have chosen, how do we know we are right?

Even in cases where there seem to be some obvious choices for the basic outcomes, it is far from clear that they should be equally likely. For example, how do we deal with a biased coin? We can take the basic outcomes to be "heads" and "tails", just as with a fair coin, but it certainly is no longer appropriate to assign each of these outcomes probability 1/2 if the coin is biased. What are the "equally likely" outcomes in that case? Even worse difficulties arise if we try to assign a probability to the event that a particular nuclear power plant will have a meltdown. What should the set of possible events be in that case, and why should they be equally likely?

In light of these problems, philosophers and probabilists have tried to find ways of viewing probability that do not depend on assigning basic events equal likelihood. Perhaps the two most common views are that (1) the numbers represent relative frequencies and (2) the numbers reflect subjective assessments of likelihood.

The intuition behind the relative-frequency interpretation is easy to explain. For example, we say that the probability that a coin lands heads is half because we expect that if we toss a coin sufficiently often, roughly half the time it will land heads. If we say that the probability that a biased coin lands heads is .6, this is because we expect that if we toss it sufficiently often, then roughly 60% of the time it will land heads. It is also easy to see that the relative-frequency interpretation satisfies the additivity property P2.

The relative-frequency interpretation is closely related to the intuition behind the principle of indifference. In the case of a coin, roughly speaking, the possible worlds now become the outcomes of a large number of coin tosses. If the coin is fair, then roughly half of the outcomes should be heads and half should be tails. If the coin is biased, the fraction of outcomes that are heads should reflect the bias. Thus, if we view all outcomes as equally probable (applying the principle of indifference), it seems reasonable to say that the probability of the coin landing heads should be roughly the fraction of outcomes in which the coin landed heads.

While this interpretation seems quite natural and intuitive, and certainly has been successfully used by the insurance industry and the gambling industry to make significant amounts of money, it also has its problems. Some of them can already be seen in the informal definition. We must toss the coin "sufficiently often". But what is sufficiently often? Is it 100 times? 1,000 times? 1,000,000 times? And what exactly does "roughly half the time" mean? Note that we cannot say "exactly half the time". If we toss the coin an odd number of times, it cannot land heads exactly half the time. And even if we toss it an even number of times, it is in fact very unlikely that it will land heads *exactly* half of those times.

To make matters worse, if we want to assign probabilities to an event like "the nuclear power plant will have a meltdown in the next five years", it is hard to think in terms of relative frequency. While it is easy to imagine tossing a coin repeatedly, it is somewhat harder to capture the sequence of events that lead to a nuclear meltdown and imagine them happening repeatedly.

Many attempts have been made to deal with these problems, perhaps the most successful being due to von Mises. It is beyond the scope of this book to discuss them. The main message that the reader should derive is that, while the intuition behind relative frequency is a very powerful one (and is certainly a compelling justification for the use of probability in some cases), it is quite difficult (some would argue impossible) to extend it to all cases where we would like to apply probability.

The relative-frequency interpretation takes probability to be an objective property of a situation. The (extreme) subjective viewpoint argues that there is no such thing as an objective notion of probability; probability is a number assigned by an individual representing his or her subjective assessment of likelihood. But if that is the case, why should these assignments of numbers obey the laws of probability?

There have been various attempts to argue that they should. The most famous of these arguments, originally due to Ramsey, is in terms of betting behavior. I discuss a variant of Ramsey's argument here. Given a set W of possible worlds and a subset $U \subseteq W$, consider an agent who can evaluate bets of the form "If U happens (that is, if the actual world is in U) then I win \$100$(1-\alpha)$ while if U doesn't happen then I lose \$100α", for $0 \leq \alpha \leq 1$. In particular, assume that when offered two such bets, she can always decide which one she prefers. (More precisely, assume that she has a total preference order on such bets; she may consider two bets equally attractive.) Let us denote such a bet as (U, α). Note that $(U, 0)$ is a "can't lose" proposition for the agent. She wins \$100 if U is the case, and loses 0 if it is not. The bet becomes less and less attractive as α gets larger; she wins less if U is the case and loses more if it is not. The worst case is if $\alpha = 1$. $(U, 1)$ is a "can't win" proposition; she wins nothing if U is true and loses \$100 if it is false. By way of contrast, the bet $(\overline{U}, 1-\alpha)$ is a can't lose proposition if $\alpha = 1$ and becomes less and less attractive as α approaches 0.

Now suppose the agent has to choose between (U, α) and $(\overline{U}, 1-\alpha)$. Which she prefers clearly depends on α. Define an agent to be *rational* if she satisfies the following two properties.

RAT1. If (U, α) is guaranteed to give more money than (U, β), then the agent prefers (U, α) to (V, β).

RAT2. Preferences are transitive, so that if (U, α) is preferred to (V, β) and (V, β) is preferred to (V', γ), then (U, α) is preferred to (V', β).

Certainly if the agent is rational, she will certainly prefer (U, α) to $(\overline{U}, 1-\alpha)$ if $\alpha = 0$ and prefer $(\overline{U}, 1-\alpha)$ if $\alpha = 1$. By transitivity, if she prefers (U, α) to $(\overline{U}, 1-\alpha)$, she also prefers (U, α') to $(\overline{U}, 1-\alpha')$ for all $\alpha' \leq \alpha$; similarly, if she prefers $(\overline{U}, 1-\beta)$ to (U, β), then she prefers $(\overline{U}, 1-\beta')$ to (U, β') for all $\beta' \geq \beta$. Let $\alpha_U = \sup\{\beta$: the agent prefers (U, β) to $(\overline{U}, 1-\beta)$. It is not hard to show that an agent satisfying RAT1 and RAT2 prefers (U, α) to $(\overline{U}, 1-\alpha)$ for all $\alpha < \alpha_U$ and prefers $(\overline{U}, 1-\alpha)$ to (U, α) for all $\alpha > \alpha_U$ (Exercise 3.2).

Intuitively, α_U is a measure of the likelihood (according to the agent) of U. The more likely she thinks U is, the higher α_U should be. If she thinks that U is certainly the case (that is, if she is certain that the actual world is in U), then α_U should be 1. For any $\alpha > 0$, she should prefer (U, α) to $(\overline{U}, 1-\alpha)$, since she feels that with (U, α), she is guaranteed to win \$100α, while with $(\overline{U}, 1-\alpha)$ she is guaranteed to lose the same amount.

Similarly, if she is certain that U is not the case, then α_U should be 0. More significantly, it can be shown that if U_1 and U_2 are disjoint sets, than a rational agent should take $\alpha_{U_1 \cup U_2} = \alpha_{U_1} + \alpha_{U_2}$. More precisely, as is shown in Exercise 3.2, if $\alpha_{U_1 \cup U_2} \neq \alpha_{U_1} + \alpha_{U_2}$, then there is a collection of bets that the agent would accept (based on her avowed preferences) according to which she is guaranteed to lose money. Such a collection of bets is called in the literature a *Dutch Book*. (Of course, this is not a literary book, but book as in "bookie" or "bookmaker".)

This discussion is summarized by the following theorem.

Theorem 3.1.3 *If an agent satisfies RAT1 and RAT2, then for each subset U of W, there is a number α_U such that the agent prefers (U, α) to $(\overline{U}, 1-\alpha)$ for all $\alpha < \alpha_U$ and prefers $(\overline{U}, 1-\alpha)$ to (U, α) for all $\alpha > \alpha_U$. Moreover, the function defined by $\mu(U) = \alpha_U$ is a probability measure.*

Proof See Exercise 3.2. ∎

This seems to be a compelling argument that if an agent's preferences can be expressed numerically, then they should obey the rules of probability. However, there is are two major assumptions hidden in such Dutch Book arguments. The argument requires that if an agent prefers (U_i, α_i) to (V_i, β_i) for each $i = 1, \ldots, k$ in isolation, then the agent prefers the collection of bets $(U_1, \alpha_1), \ldots, (U_k, \alpha_k)$ to the collection $(V_1, \beta_1), \ldots, (V_k, \beta_k)$. While this is certainly a natural assumption, it is by no means vacuous. A rational agent may certainly prefer each of (U_i, α_i) to (V_i, β_i) in isolation but react differently to the package. This, in fact, is a large part of the theory behind hedge funds. A second implicit assumption is that the agent can always tell which of the two options she prefers. One could instead imagine an agent that had numbers $\alpha_1 < \alpha_2$ such that she would prefer (U, α) to $(\overline{U}, 1 - \alpha)$ for $\alpha < \alpha_1$ and prefer $(\overline{U}, 1 - \alpha)$ to (U, α) for $\alpha > \alpha_2$, but in the interval between α_1 and α_2, wasn't sure which was better. This certainly doesn't seem so irrational. These assumptions make Dutch Book arguments somewhat less compelling than they might at first appear.

It might also seem worrisome that the subjective probability interpretation puts no constraints on the agent's subjective likelihood other than the requirement that it obey the laws of probability. In the case of tossing a fair die, for example, taking each outcome to be equally likely seems "right". It may seem unreasonable for someone who subscribes to the subjective point of view to be able to put probability .8 on the die landing 1, and probability .04 on each of the other five possible outcomes. More generally, when it seems that the principle of indifference is applicable or if we have detailed frequency information, should the subjective probability take this into account? The standard responses to this concern are (1) indeed frequency information and the principle of indifference should be taken into account, when appropriate and (2) even if they are not taken into account, all choices of initial subjective probability will eventually converge to the same probability measure as more information is received; the measure that they converge to will in some sense be the "right" one (see Example 4.2.2).

Different readers will probably have different feelings as to how compelling these and other defenses of probability really are. However, the fact that philosophers have come up with a number of independent justifications for probability is certainly a strong point in its favor. Much more effort has gone into justifying probability than any other approach for representing uncertainty. Time will tell if equally compelling justifications can be given for other approaches. In any case, there is no question that probability is the most widely accepted and widely used approach to representing uncertainty today.

3.2 Lower and Upper Probabilities

Despite its widespread acceptance, there have been many criticisms of probability. Three of the most common ones are (1) probability is not good at representing ignorance, (2) while an agent may be prepared to assign probabilities to some sets, she may not be prepared to assign probabilities to all sets, and (3) while an agent may be willing in principle to assign probabilities to all the sets in some algebra, computing these probabilities requires some computational effort; she may simply not have the computational resources required to do it. As we shall see, these criticisms turn out to be closely related to one of the criticisms of the Dutch Book justification for probability mentioned in Section 3.1.1. The following two examples might help clarify the issues.

Example 3.2.1 Suppose that a coin that is tossed once. There are two possible worlds, *heads* and *tails*, corresponding to the two possible outcomes. If the coin is known to be fair, it seems reasonable to assign probability 1/2 to each of these worlds. However, suppose that the coin has an unknown bias. How do we represent our ignorance in this case? We might say that, given that we have no idea how the coin will land, we still view both *heads* and *tails* as equally likely. However, there seems to be a significant qualitative difference between the two situations. Is there some way we can capture the difference? One possibility is to take the bias of the coin to be part of the possible world (that is, a possible world describes both the bias of the coin and the outcome of the toss), but then what is the probability of *heads*? ∎

Example 3.2.2 Suppose that we have a bag of 100 marbles; 30 are known to be red, and the remainder are known to be either blue or yellow, although the exact proportion of blue and yellow is not known. We pick a marble out of the bag. We can model this with three possible worlds, *red*, *blue*, and *yellow*, one for each of the possible outcomes. It seems reasonable to assign probability .3 to the outcome to choosing a red marble, and thus probability .7 to choosing either blue or yellow, but what probability do we assign to the other two outcomes? ∎

There are two related ways of handling these problems. One is to simply represent our ignorance not with one probability measure, but with a set of them. For example, in the case of the coin with unknown bias, we can use the set $\mathcal{P}_1 = \{\mu_a : a \in [0,1]\}$ of probability measures, where μ_a gives *heads* probability a. Similarly, in the case of the marbles, we can use the set $\mathcal{P}_2 = \{\mu'_a : a \in [0,.7]\}$, where μ'_a gives *red* probability .3, *blue* probability a, and *yellow* probability $.7 - a$.

An alternative approach is to make only some sets measurable. Intuitively, the measurable sets are the ones to which we are prepared to assign a probability. For example, in the case of the coin, we could take our algebra to consist only of the empty set and $\{heads, tails\}$, so that $\{heads\}$ and $\{tails\}$ are no longer measurable sets. Clearly, there is only one probability measure on this space; for future reference, call it μ_1. Thus, by considering this trivial algebra, we avoid having to consider what probability to assign to $\{heads\}$ or $\{tails\}$.

Similarly, in the case of the marbles, we can consider the algebra

$$\{\emptyset, \{red\}, \{blue, yellow\}, \{red, yellow, blue\}\}.$$

There is an obvious probability measure μ_2 on this algebra that describes the story in Example 3.2.2: we simply take $\mu_2(red) = .3$. This determines all the other probabilities.

Notice that, with the first approach, in the case of the marbles, we can say that the probability of *red* is .3 (since all probability measures $\mathcal{P}_2$ give *red* probability .3), but all we can say about the probability of *blue* is that it is somewhere between 0 and .7 (since that is the range of possible probabilities for *blue* according to the probability measures in $\mathcal{P}_2$), and similarly for *yellow*. There is a sense in which the second approach also gives this answer: any probability for *blue* between 0 and .7 is compatible with the probability measure μ_2. Similarly, in the case of the coin with an unknown bias, all we can say about the probability of *heads* is that it is somewhere between 0 and 1.

If we were to recast these examples in terms of Dutch book argument, the fact that, for example, all we can say about the probability of the marble being blue is that it is between 0 and .7 corresponds to the agent definitely preferring $(\overline{blue}, 1-\alpha)$ to $(blue, \alpha)$ for $\alpha > .7$, but not being able to choose between the two bets for $0 \le \alpha \le .7$. We can in fact recast the Dutch book arguments for probability to provide a justification for using sets of probabilities, once we drop the assumption that the agent can always decide which of (U, α) and $(\overline{U}, 1-\alpha)$ she prefers, for arbitrary U and α.

Notice that, with both approaches, all that we can say about the probability of *blue* is that it is between 0 and .7. The fact that the two approaches give the same answer is no accident. They are in fact closely related, as I now show.

Given a set $\mathcal{P}$ of probability measures on a set W of possible worlds and $U \subseteq W$, define

$$\begin{aligned}\mathcal{P}_*(U) &= \inf\{\mu(U) : \mu \in \mathcal{P}\}\\ \mathcal{P}^*(U) &= \sup\{\mu(U) : \mu \in \mathcal{P}\},\end{aligned}$$

where, as usual, sup denotes “least upper bound” and inf denotes “greatest lower bound”. $\mathcal{P}_*(U)$ is called the *lower probability* of U and $\mathcal{P}^*(U)$ is called the *upper probability* of U. For example, $(\mathcal{P}_2)_*(blue) = 0$, $(\mathcal{P}_2)^*(blue) = .7$, and similarly for *yellow*, while $(\mathcal{P}_2)_*(red) = (\mathcal{P}_2)^*(red) = .3$.

Now consider the approach of taking only some subsets to be measurable. An algebra $\mathcal{F}'$ is a *subalgebra* of an algebra $\mathcal{F}$ if $\mathcal{F}' \subseteq \mathcal{F}$. If $\mathcal{F}'$ is a subalgebra of $\mathcal{F}$, μ' is a probability measure on $\mathcal{F}'$, and μ is a probablity measure on $\mathcal{F}$, then μ is an *extension* of μ' if μ' and μ agree on all sets in $\mathcal{F}'$. Notice that $\mathcal{P}_1$ consists of all the extensions of μ_1 to the algebra consisting of all subsets of $\{heads, tails\}$ and $\mathcal{P}_2$ consists of all extensions of μ_2 to the algebra of all subsets of $\{red, blue, yellow\}$.

If μ is a probability measure on the subalgebra $\mathcal{F}'$ and $U \in \mathcal{F} - \mathcal{F}'$, then $\mu(U)$ is undefined, since U is not in the domain of μ. We can extend μ to $\mathcal{F}$ in two standard ways, by defining function μ_* and μ^*, traditionally called the *inner measure* and *outer measure induced by* μ, respectively. For $U \in \mathcal{F}$, we define:

$$\begin{aligned}\mu_*(U) &= \sup\{\mu(V) : V \subseteq U, V \in \mathcal{F}'\}\\ \mu^*(U) &= \inf\{\mu(V) : V \supseteq U, V \in \mathcal{F}'\}.\end{aligned}$$

These definitions are perhaps best understood in the case that the set of possible worlds (and hence the algebra $\mathcal{F}'$) is finite. In that case, μ_* is the measure of the largest measurable set (in $\mathcal{F}'$) contained in U, and μ^* is the measure of the smallest measurable set containing U. That is, $\mu_*(U) = \mu(V_1)$, where $V_1 = \cup_{B \subseteq U,\, B \in \mathcal{F}} B$ and $\mu^*(U) = \mu(V_2)$, where $V_2 = \cap_{U \subseteq B,\, B \in \mathcal{F}} B$ (Exercise 3.3). Intuitively, $\mu_*(U)$ is the best approximation to the actual probability of U from below and $\mu^*(U)$ is the best approximation from above. If $U \in \mathcal{F}'$, then it is easy to see that $\mu_*(U) = \mu^*(U) = \mu(U)$. If $U \in \mathcal{F} - \mathcal{F}'$, then in general we have $\mu_*(U) < \mu^*(U)$.

For example, $(\mu_2)_*(blue) = 0$ and $(\mu_2)^*(blue) = .7$, since the largest measurable set contained in $\{blue\}$ is the empty set, while the smallest measurable set containing *blue* is $\{blue, yellow\}$. In addition, $(\mu_2)_*(red) = (\mu_2)^*(red) = \mu_2(red) = .3$. These are precisely the same numbers obtained using the lower and upper probabilities $(\mathcal{P}_2)_*$ and $(\mathcal{P}_2)^*$. Of course, this is no accident.

Theorem 3.2.3 *Let μ be a probability measure on a subalgebra $\mathcal{F}' \subseteq \mathcal{F}$ and let $\mathcal{P}_\mu$ consist of all extensions of μ to $\mathcal{F}$. Then $\mu_*(U) = (\mathcal{P}_\mu) * (U)$ and $\mu^*(U) = (\mathcal{P}_\mu)^*(U)$ for all $U \in \mathcal{F}$.*

Proof See Exercise 3.4. ∎

Note that whereas probability measures are additive, so that if U and V are disjoint sets, we have $\mu(U \cup V) = \mu(U) + \mu(V)$, inner measures are *superadditive* and outer measures are *subadditive*, so that

$$\begin{array}{l} \mu_*(U \cup V) \geq \mu_*(U) + \mu_*(V) \\ \mu^*(U \cup V) \leq \mu^*(U) + \mu^*(V). \end{array} \tag{3.1}$$

In addition, the relationship between inner and outer measures is defined by

$$\mu_*(U) = 1 - \mu^*(\overline{U}) \tag{3.2}$$

(Exercise 3.5).

In fact, we can say even more. Probability measures satisfy what has been called the *inclusion-exclusion* rule, which describes how to compute the probability of (not necessarily disjoint) sets. In the case of two sets, this rule is

$$\mu(U \cup V) = \mu(U) + \mu(V) - \mu(U \cap V). \tag{3.3}$$

To see this, note that $U \cup V$ can be written as the union of three disjoint sets, $U - V$, $V - U$, and $U \cap V$. Thus,

$$\mu(U \cup V) = \mu(U - V) + \mu(V - U) + \mu(U \cap V).$$

Since U is the union of $U - V$ and $U \cap V$ and V is the union of $V - U$ and $U \cap V$, we get

$$\begin{array}{l} \mu(U) = \mu(U - V) + \mu(U \cap V) \\ \mu(V) = \mu(V - U) + \mu(U \cap V). \end{array}$$

Now Equation (3.3) easily follows by simple algebra.

Intuitively, what is going on here is that if we try to compute $\mu(U \cup V)$ by adding $\mu(U)$ and $\mu(V)$, we are counting $\mu(U \cap V)$ twice, so we must subtract it once. If U and V are disjoint, then $\mu(U \cap V) = 0$, and we get additivity as a special case of the inclusion-exclusion rule in the case of two sets. In the case of three sets U, V, W, illustrated in Figure **??**, similar arguments show that

$$\mu(U \cup V \cup W) = \mu(U) + \mu(V) + \mu(W) - \mu(U \cap V) - \mu(U \cap W) - \mu(V \cap W) + \mu(U \cap V \cap W). \tag{3.4}$$

That is, we add the sets (these are one-way intersections), subtract the two-way intersections, and add the three-way intersections. More generally, we have the full-blown inclusion-exclusion rule:

$$\mu(\cup_{i=1}^n U_i) = \sum_{i=1}^n \sum_{\{I:|I|=i\}} (-1)^{i+1}\mu(\cap_{j\in I} U_j). \tag{3.5}$$

Equation (3.5) says that to compute the probability of the union of n sets, we add the probability of each of the sets (the case when $|I| = 1$), subtract the probability of the two-way intersections (the case when $|I| = 2$), add the probability of the three-way intersections, and so on. (The $(-1)^{i+1}$ term ensures that we switch from addition to subtraction and back again as the size of the intersection set increases.) Equations (3.3) and (3.4) are just special cases of the general rule when $n = 2$ and $n = 3$. I leave it to the reader to verify this rule (Exercise 3.6).

Now for inner measures and outer measures, we also have an inclusion-exclusion rule, except that we replace $=$ by $\geq$ in the case of inner measures and by $\leq$ in the case of outer measures. Thus, for example,

$$\mu_*(\cup_{i=1}^n U_i) \geq \sum_{i=1}^n \sum_{\{I:|I|=i\}} (-1)^{i+1}\mu_*(\cap_{j\in I} U_j) \tag{3.6}$$

(Exercise 3.8), and similarly for μ^*, except that $\geq$ is replaced by $\leq$. As we shall see in Section **??**, there is a sense in which these inequalities characterize inner and outer measures.

Using sets of probability measures and computing lower and upper probabilities or using a single probability measure defined on an algebra that does not necessarily include all events of interest and computing inner and outer measures are two ways in which we can deal with imprecision and ignorance. Sets of probability measures are more general, in that they can capture more situations. For example, suppose that we know not only that there are 70 blue and yellow marbles altogether, but that the difference between the number of blue and yellow marbles is at least 10 (that is, if there are b blue marbles and y yellow marbles, that $|b - y| \geq 10$). In this case, it seems appropriate to consider the set $\mathcal{P}_2'$ consisting of all probability measures μ on $\{red, yellow, blue\}$ such that $\mu(red) = .3$ and $|\mu(blue) - \mu(yellow)| \geq .1$. This set of measures is not the set of all extensions of any measure on the algebra $\mathcal{F}'$ generated by $\{red\}$ and $\{blue, yellow\}$. We still have $(\mathcal{P}_2')_*(blue) = 0$ and $(\mathcal{P}_2')^*(blue) = .7$, and similarly for *yellow*, just as in the case of $\mathcal{P}_2$, but there is extra information in $\mathcal{P}_2'$ that is not being captured by the lower and upper probabilities.

Because sets of probability measures are more general than probability measures defined on a subalgebra, they satisfy fewer properties. They do satisfy analogues of Equations (3.1) and (3.2) (with μ_* and μ^* replaced by $\mathcal{P}_*$ and $\mathcal{P}^*$, respectively) but they do not satisfy the analogue of Equation (3.6) in general (Exercise 3.9).

3.3 Dempster-Shafer Belief Functions

The Dempster-Shafer theory of evidence, originally introduced by Arthur Dempster and then developed by Glenn Shafer, provides another approach to attaching likelihood to events. This approach starts out with a *belief function* (sometimes called a *support function*). Given a set W of possible worlds and $U \subseteq W$, the belief in U, denoted $\mathrm{Bel}(U)$ is a number in the interval $[0,1]$. (Thus, we can think of Bel as being defined on the algebra 2^W, consisting of all subsets of W. The definition can be easily generalized so that the domain of Bel is some algebra over W, although this is typically not done in the literature.) We can think of $\mathrm{Bel}(U)$ as providing a lower bound on the likelihood of U. There is a corresponding number $\mathrm{Plaus}(U)$ called the *plausibility* of U that places an upper bound on the likelihood of U. Thus, to every event U, we can attach the interval $[\mathrm{Bel}(U), \mathrm{Plaus}(U)]$. We assume that a belief function Bel defined on a space W satisfies the following three properties.

B1. $\mathrm{Bel}(\emptyset) = 0$.

B2. $\mathrm{Bel}(W) = 1$.

B3. $\mathrm{Bel}(\cup_{i=1}^n U_i) \geq \sum_{i=1}^n \sum_{\{I:|I|=i\}} (-1)^{i+1} \mathrm{Bel}(\cap_{j\in I} U_j)$.

B1 and B2 are what we would expect, and hold for probability as well. B3 is just the inclusion-exclusion rule with $=$ replaced by $\geq$. Thus, every probability measure is a belief function. Moreover, from the results of the previous section, it follows that every inner measure is a belief function too. The converse does not hold; that is, not every belief function is an inner measure corresponding to some probability measure. For example, if $W = \{w, w'\}$, $\mathrm{Bel}(w) = 1/2$, $\mathrm{Bel}(w') = 0$, $\mathrm{Bel}(W) = 1$, and $\mathrm{Bel}(\emptyset) = 0$, then it is easy to see that Bel is a belief function, but there is no probability measure μ on W such that $\mathrm{Bel} = \mu_*$ (Exercise 3.10). On the other hand, Exercise 5.4 shows that there is a sense in which every belief function can be identified with the inner measure corresponding to some probability measure.

As we observed, a probability measure defined on 2^W can be characterized by its behavior on singleton sets. This is not the case for belief functions. For example, it is easy to construct two belief functions Bel_1 and Bel_2 on $\{1,2,3\}$ such $\mathrm{Bel}_1(i) = \mathrm{Bel}_2(i) = 0$ for $i = 1,2,3$ (so that Bel_1 and Bel_2 agree on singleton sets) but $\mathrm{Bel}_1(\{1,2\}) \neq \mathrm{Bel}_2(\{1,2\})$ (Exercise 3.11).

Define $\mathrm{Plaus}(U) = 1 - \mathrm{Bel}(\overline{U})$. A plausibility function bears the same relationship to a belief function that an outer measure bears to an inner measure. Indeed, every outer measure is a plausibility function. It follows easily from B3 (applied to U and $\overline{U}$, with $n = 2$) that $\mathrm{Bel}(U) \leq \mathrm{Plaus}(U)$ (Exercise 3.12). Thus, to every event U, we can attach the interval $[\mathrm{Bel}(U), \mathrm{Plaus}(U)]$.

These observations show that there is a close relationship between belief functions, inner measures, and lower probabilities. Part of this relationship is made precise by the following theorem.

Theorem 3.3.1 *Given a belief function* Bel *defined on a space W, there is a set $\mathcal{P}_{\mathrm{Bel}}$ of probability measures on W (with domain 2^W) such that* $\mathrm{Bel} = (\mathcal{P}_{\mathrm{Bel}})_*$ *and* $\mathrm{Plaus} = (\mathcal{P}_{\mathrm{Bel}})^*$.

Proof A probability measure μ is *consistent with* Bel if $\mu(U) \geq \mathrm{Bel}(U)$ for every subset $U \subseteq W$. Let $\mathcal{P}_{\mathrm{Bel}}$ consist of all probability measures consistent with Bel. The proof that $\mathcal{P}_{\mathrm{Bel}}$ has the required properties is left to the reader (Exercise 3.13). ∎

Theorem 3.3.1 shows that every belief function on W can be viewed as a lower probability of a set of probability measures on W. The converse is not true in general. It follows from Exercise 3.9 that lower probabilities do not necessarily satisfy the analogue of Equation (3.6), and thus there is a space W and a set $\mathcal{P}$ of probability measures on W such that there is no belief function Bel on W with $\mathrm{Bel} = \mathcal{P}_*$. (In fact, we can take $W = \{1,2,3\}$ and take $\mathcal{P}$ to be a set of two probability measures.)

In any case, while belief functions can be understood (to some extent) in terms of lower probability, this is not the only way of understanding them. Belief functions are part of a theory of *evidence*. The intuition behind this theory is that we get evidence that supports events to varying degrees. For example, in the case of the marbles, the information that there are exactly 30 red marbles provides support in degree .3 for *red*; the information that there are 70 yellow and blue marbles does not provide any positive support for either *blue* or *yellow*, but does provide support .7 for $\{$*blue*, *yellow*$\}$. In general, the picture is that evidence provides some degree of support (possibly 0) for each subset of W. The total amount of support is 1. The belief that U holds, $\mathrm{Bel}(U)$, is then the sum of all of the support on subsets of U.

Formally, this is captured as follows. A *mass function* (sometimes called a *basic probability assignment*) on W is a function $m : 2^W \to [0,1]$ satisfying the following properties.

M1. $m(\emptyset) = 0$.

M2. $\sum_{U \subseteq W} m(U) = 1$.

Given a mass function m, define the belief function based on m, Bel_m, by taking

$$\mathrm{Bel}_m(U) = \sum_{\{U' : U' \subseteq U\}} m(U'). \tag{3.7}$$

The corresponding plausibility function Plaus_m is defined as

$$\mathrm{Plaus}_m(U) = \sum_{\{U' : U' \cap U \neq \emptyset\}} m(U').$$

It is obvious that Bel_m satisfies B1 and B2. It is not as obvious, but is nonetheless true, that Bel_m satisfies B3, and is thus in fact a belief function. Moreover, the following theorem shows that every belief function arises from some mass function m.

Theorem 3.3.2 *Given a mass function m of a finite set W, the function Bel_m is a belief function and Plaus_m is the corresponding plausibility function. Moreover, given a belief function Bel on W, there is a unique mass function m on W such that $\mathrm{Bel} = \mathrm{Bel}_m$.*

Proof See Exercise 3.14. ∎

Bel_m and its corresponding plausibility function Plaus_m are the belief function and plausibility function *corresponding to* the mass function m.

Example 3.2.2 can be captured using the function m such that $m(\mathit{red}) = .3$, $m(\mathit{blue}) = m(\mathit{yellow}) = m(\{\mathit{red}, \mathit{blue}, \mathit{yellow}\}) = 0$, and $m(\{\mathit{blue}, \mathit{yellow}\}) = .7$. In this case, m looks like a probability measure, since the sets that get positive mass are disjoint, and the masses sum to 1. However, in general, the sets of positive mass may not be disjoint. It is perhaps best to think of $m(U)$ as the amount of belief committed to U that has not already been committed to its subsets. The following example should help make this clear

Example 3.3.3 Suppose a physician sees a case of jaundice. He considers four possible hypotheses regarding its cause: hepatitis (*hep*), cirrhosis (*cirr*), gallstone (*gall*), and pancreatic cancer (*pan*). For simplicity, suppose that these are the only causes of jaundice, and that a patient with jaundice suffers from exactly one of these problems. Thus, we can take the set W of possible worlds to be $\{hep, cirr, gall, pan\}$. Only some subsets of 2^W are of diagnostic significance. There are tests whose outcomes support each of the individual hypotheses, and tests that support *intrahepatic cholestasis*–$\{hep, cirr\}$–and *extrahepatic cholestasis*–$\{gall, pan\}$; the latter two tests do not provide further support for the individual hypotheses.

If there is no information supporting any of the hypotheses, this would be represented by a mass function that assigns mass 1 to W and mass 0 to all other subsets of W. On the other hand, suppose there is evidence that supports intrahepatic cholestasis to degree .7. (The degree to which evidence supports a subset of W can be given both a relative frequency and a subjective interpretation. Under the relative frequency interpretation, it could be the case that 70% of the time that the test had this outcome, a patient had hepatitis or cirrhosis.) We would represent this by the mass function that assigns .7 to $\{hep, cirr\}$ and the remaining .3 to W. The fact that the test provides support only .7 to $\{hep, cirr\}$ does not mean that it provides support .3 for its complement, $\{gall, pan\}$. Rather, the remaining .3 is viewed as uncommitted. As a result, $\mathrm{Bel}(\{hep, cirr\} = .7$ and $\mathrm{Plaus}(\{hep, cirr\}) = 1$. ∎

Suppose a doctor performs two tests on a patient, each of which provide some degree of support for a particular hypothesis. We would like some way of combining the evidence; the Dempster-Shafer theory provides a way of doing this.

Let Bel_1 and Bel_2 denote two belief functions on some set W, and let m_1 and m_2 be their respective mass functions. *Dempster's Rule of Combination* provides a way of constructing a new mass function $m_1 \oplus m_2$, provided that there are at least two sets U_1 and U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. If there are no such sets U_1 and U_2, then $m_1 \oplus m_2$ is undefined. Notice that, in this case, there must be disjoint sets V_1 and V_2 such that $\mathrm{Bel}_1(V_1) = \mathrm{Bel}_2(V_2) = 1$ (Exercise 3.15). Thus, Bel_1 and Bel_2 describe diametrically opposed beliefs, so it should come as no surprise that they cannot be combined.

To understand the Rule of Combination, note that, by simple algebra and M2, we have

$$1 = (\sum_{U_1 \subseteq W} m_1(U_1))(\sum_{U_2 \subseteq W} m_2(U_2)) = \sum_{U_1, U_2 \subseteq W} m_1(U_1)m_2(U_2).$$

Roughly speaking, $m_1 \oplus m_2$ apportions mass $m_1(U_1)m_2(U_2)$ to $U_1 \cap U_2$. But there are in general many ways of writing a set U as an intersection $U_1 \cap U_2$; the mass that $m_1 \oplus m_2$ assigns to a set U is the sum of $m_1(U_1)m_2(U_2)$ for all ways for writing U as $U_1 \cap U_2$.

There is only one problem with this. If $U_1 \cap U_2 = \emptyset$, then by M1, we must assign $U_1 \cap U_2$ mass 0. Thus, we can apply the intuition above only as long as $U_1 \cap U_2 \neq \emptyset$. Then to make sure that M2 is satisfied, we must renormalize. Define $(m_1 \oplus m_2)(\emptyset) = 0$ and for $U \neq \emptyset$, define

$$(m_1 \oplus m_2)(U) = \sum_{\{U_1, U_2 : U_1 \cap U_2 = U\}} m_1(U_1)m_2(U_2)/c,$$

where $c = \sum_{\{U_1, U_2 : U_1 \cap U_2 \neq \emptyset\}} m_1(U_1)m_2(U_2)$. If $m_1 \oplus m_2$ is defined, then $c > 0$, since there are sets U_1, U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. The choice of c guarantees that $m_1 \oplus m_2$ satisfies M2. Let $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ be the belief function corresponding to $m_1 \oplus m_2$.

The Dempster rule of combination is claimed to be appropriate when combining two *independent* pieces of evidence. Independence is to be viewed as an intuitive, primitive notion here. Essentially, it says the sources of the evidence are unrelated. The rule has the attractive feature of being commutative and associative:

$$m_1 \oplus m_2 = m_2 \oplus m_1 \text{ and}$$

$$m_1 \oplus (m_2 \oplus m_3) = (m_1 \oplus m_2) \oplus m_3.$$

This seems reasonable: our final beliefs should be independent of the order and the way in which we combine the evidence. Let m_{vac} be the *vacuous mass function* on W: $m_{vac}(W) = 1$ and $m(U) = 0$ for $U \subset W$. It is easy to check that m_{vac} is the neutral element in the space of mass functions on W: For any mass function m, we have $m_{vac} \oplus m = m \oplus m_{vac} = m$ (Exercise 3.16).

Rather than going through a formal derivation of the Rule of Combination, I consider two examples of its use here, where it gives intuitively reasonable results. In Section 4.2, it is related to the probabilistic combination of evidence.

Example 3.3.4 Returning to the medical situation in Example 3.3.3, suppose that two tests are carried out. The first confirms hepatitis to degree .8 and says nothing about the other hypotheses; we capture this by the mass function m_1 such that $m_1(\mathit{hep}) = .8$ and $m_1(W) = .2$. The second test confirms intrahepatic cholestasis to degree .6; this is captured by the mass function m_2 such that $m_2(\{\mathit{hep}, \mathit{cirr}\}) = .6$ and $m_2(W) = .4$. A straightforward computation shows that

$$\begin{aligned}
&(m_1 \oplus m_2)(\mathit{hep}) = .8,\\
&(m_1 \oplus m_2)(\{\mathit{hep}, \mathit{cirr}\}) = .12,\\
&(m_1 \oplus m_2)(W) = .08. \blacksquare
\end{aligned}$$

Example 3.3.5 A coin is said to have *bias* α if it lands heads with probability α and tails with probability $1-\alpha$. Suppose that Alice has a coin and she knows that it either has bias 2/3 (BH) or bias 1/3 (BT). Initially, she has no evidence for BH or BT. This is captured by the vacuous belief function Bel_{init}, where $\mathrm{Bel}_{init}(BH) = \mathrm{Bel}_{init}(BT) = 0$ and $\mathrm{Bel}_{init}(W) = 1$. Suppose Alice then tosses the coin and observes that it lands heads. This should give her some positive evidence for BH but no evidence for BT. One way to capture this evidence is by using the belief function Bel_{heads} such that $\mathrm{Bel}_{heads}(BH) = \alpha > 0$ and $\mathrm{Bel}_{heads}(BT) = 0$. (The exact choice of α does not matter.) The corresponding mass function m_{heads} is such that $m_{heads}(BT) = 0$, $m_{heads}(BH) = \alpha$ and $m_{heads}(W) = 1-\alpha$. We can similarly define mass and belief functions m_{tails} and Bel_{tails} that capture the evidence of tossing the coin and seeing tails. Note that $m_{init} \oplus m_{heads} = m_{heads}$, and similarly for m_{tails}. Combining Alice's initial ignorance regarding BH and BT with the evidence results in the same beliefs as those produced by just the evidence itself.

Now what happens if Alice observes k heads in a row? Intuitively, this should increase her degree of belief that the coin is biased towards heads. Let $m^k_{heads} = m_{heads} \oplus \cdots \oplus m_{heads}$ (k times). A straightforward computation shows that $m^k_{heads}(BT) = 0$, $m^k_{heads}(BH) = 1-(1-\alpha)^k$, and $m^k_{heads}(W) = \alpha^k$. As we would expect, observing heads more and more often drives Alice's belief that the coin is biased towards heads to 1.

Another straightforward computation shows that

$$m_{heads} \oplus m_{tails}(BH) = m_{heads} \oplus m_{tails}(BT) = \alpha(1-\alpha)/(1-\alpha^2).$$

Thus, as would be expected, after seeing heads and then tails (or, since $\oplus$ is commutative, after seeing tails and then heads), Alice assigns an equal degree of belief to BH and BT. However, unlike the initial situation where Alice assigned no belief to either BH or BT, she now assigns positive belief to each of them, since she has seen some evidence in favor of each. ∎

3.4 Possibility Measures

Possibility measures are yet another approach to assigning numbers to sets. They are based on ideas of *fuzzy logic*. The basic idea underlying possibility measures is easy to explain. Suppose for simplicity that we are in the finite case, and all sets are measurable. A *possibility measure* Poss associates with each subset of W a number in $[0,1]$ and satisfies the following three properties.

Poss1. $\mathrm{Poss}(\emptyset) = 0$.

Poss2. $\text{Poss}(W) = 1$.

Poss3. $\text{Poss}(U \cup V) = \max(\text{Poss}(U), \text{Poss}(V))$ if U and V are disjoint.

The only difference between probability and possibility is that if A and B are disjoint sets, then $\text{Poss}(U \cup V)$ is the maximum of $\text{Poss}(U)$ and $\text{Poss}(V)$, while $\mu(U \cup V)$ is the sum of $\mu(U)$ and $\mu(V)$. (It is easy to see that Poss3 holds even if U and V are not disjoint; see Exercise 3.17. By way of contrast, P2 does not hold if U and V are not disjoint.)

It follows that, like probability, in a finite space we can characterize possibility by its behavior on singleton sets; $\text{Poss}(U) = \max_{u \in U} \text{Poss}(u)$. For Poss2 to be true, it must be the case that $\max_{w \in W} \text{Poss}(w) = 1$; that is, at least one element in W must have maximum possibility.

There is a dual to possibility called *necessity*. The necessity of U, $\text{Nec}(U)$, is defined as $1 - \text{Poss}(\overline{U})$. It is not hard to show that $\text{Nec}(U) \leq \text{Poss}(U)$ (Exercise 3.18). The relationship between necessity and possibility is reminiscent of the relationship between belief and plausibility. It turns out that possibility measures are in fact a special case of Dempster-Shafer plausibility functions. Define a mass function m to be *consonant* if it does not assign positive mass to disjoint sets. More precisely, m is a consonant mass function if $m(U) > 0$ and $m(U') > 0$ implies that either $U \subseteq U'$ or $U' \subseteq U$. The following theorem shows that possibility measures are Dempster-Shafer plausibility functions corresponding to a consonant mass function.

Theorem 3.4.1 *If m is a consonant mass function on a finite space W, then* Plaus_m*, the plausibility function corresponding to m, is a possibility measure. Conversely, given a possibility measure Poss on W, there is a consonant mass function m such that Poss is the plausibility measure corresponding to m.*

Proof See Exercise 3.19. ∎

Although we can understand possibility measures in terms of the Dempster-Shafer approach, this is perhaps not the best way of viewing them. Why restrict to consonant mass functions, for example? Many other interpretations of possibility measures have been provided, for example, in terms of degree of surprise (see the next section) and betting behavior. Perhaps the most common interpretation given to possibility and necessity is that they capture, not a degree of likelihood, but a (subjective) degree of uncertainty regarding the truth of a statement. This is viewed as being particularly appropriate for vague statements such as "John is tall". When deciding on the degree of uncertainty appropriate to such a statement, there are two issues to consider. First, we might be uncertain about John's actual height. But even if we knew that John was 1.78 meters tall (about 5 foot 10 inches), we might still be uncertain about the truth of the statement. Putting the two sources of uncertainty together, we might decide that we believe the statement to be true to degree at least .3 and at most .7. In this case, we can take the necessity of the statement to be .3 and its possibility to be .7.

Possibility measures have an important computational advantage over probability: they are compositional. Suppose that we are given arbitrary sets U and V together with $\mu(U)$ and $\mu(V)$. What can we say about $\mu(U \cup V)$? The best we can do is to provide bounds: $\mu(U \cup V)$ is certainly at least $\max(\mu(U), \mu(V)$ and it is at most $\min(\mu(U) + \mu(V), 1)$. These, in fact, are the best bounds for $\mu(U \cup V)$ in terms of $\mu(U)$ and $\mu(V)$ (Exercise 3.20). On the other hand, as Exercise 3.17 shows, we can easily determine $\mathrm{Poss}(U \cup V)$ in terms of $\mathrm{Poss}(U)$ and $\mathrm{Poss}(V)$: it is just the maximum of the two.

Of course, the question still remains as to why max is the appropriate operation for ascribing uncertainty to the union of two sets. There have been various justifications given for taking max, but a discussion of this issue is beyond the scope of the book. Interestingly, the use of max is compatible with the notion of relative likelihood defined in Section 2.3. More precisely, define $w \succeq w'$ if $\mathrm{Poss}(w) \geq \mathrm{Poss}(w')$. Then if $\succeq^s$ and $\succ^s$ are defined as in Section 2.3, it is easy to see that if $\mathrm{Poss}(w) > 0$ for all $w \in W$, then $U \succeq^s V$ iff $\mathrm{Poss}(U) \geq \mathrm{Poss}(V)$ and $U \succ^s V$ iff $\mathrm{Poss}(U) > \mathrm{Poss}(V)$ (Exercise 3.21). (Note that since $\succeq$ is a total order, $\succ^s$ and $\succ'$ agree in this case.) It follows from Proposition 2.3.4 that Poss is qualitative; that is, if $\mathrm{Poss}(U_1 \cup U_2) > \mathrm{Poss}(U_3)$ and $\mathrm{Poss}(U_1 \cup U_3) > \mathrm{Poss}(U_2)$, then $\mathrm{Poss}(U_1) > \mathrm{Poss}(U_2 \cup U_3)$; it is actually not hard to prove this directly as well (Exercise 3.22).

3.5 Ranking Functions

Another approach to representing uncertainty, somewhat similar in spirit to possibility measures, is given by what are called *(ordinal) ranking functions.* I consider a slightly simplified version here. A ranking function κ again assigns to every set a number, but this time the number is a natural number or infinity; that is, $\kappa : 2^W \rightarrow I\!N^* = I\!N \cup \{\infty\}$. ($I\!N$ denotes the natural numbers, $\{0, 1, 2, \ldots\}$.) The numbers can be thought of as denoting degrees of surprise; that is, $\kappa(U)$ is the degree of surprise the agent would feel if the actual world were in U. The higher the number, the greater the degree of surprise. 0 denotes "unsurprising", 1 denotes "somewhat surprising", 2 denotes "quite surprising", and so on; ∞ denotes "so surprising as to be impossible". For example, in a coin toss, we might assign $\kappa(\mathit{heads}) = \kappa(\mathit{tails}) = 0$ and $\kappa(\mathit{edge}) = 3$, where *edge* is the event that the coin lands on edge.

Given this intuition, it should not be surprising that ranking functions are required to satisfy the following three properties:

R1. $\kappa(\emptyset) = \infty$.

R2. $\kappa(W) = 0$.

R3. $\kappa(U \cup V) = \min(\kappa(U), \kappa(V)$ if U and V are disjoint.

(Again, R3 holds even if U and V are not disjoint; see Exercise 3.17.) Thus, with ranking functions, ∞ and 0 play the role played by 0 and 1 in probability and possibility, and min plays the role of + in probability and max in possibility. As with probability and possibility, a ranking function is characterized by its behavior on singletons; $\kappa(U) = \min_{u \in U} \kappa(u)$. To ensure that R2 holds, we require that $\min_{w \in W} \kappa(w) = 0$; that is, at least one element in W must have a rank of 0.

Ranking functions as defined here are can in fact be viewed as possibility measures in a straightforward way. Given a ranking function κ, define the possibility measure Poss_κ by taking $\mathrm{Poss}_\kappa(U) = 1/(1 + \kappa(U))$. ($\mathrm{Poss}_\kappa(U) = 0$ if $\kappa(U) = \infty$.) It is easy to see that Poss_κ is indeed a possibility measure (Exercise 3.23). This suggests that we can give possibility measures a degree-of-surprise interpretation similar in spirit to that given to ranking functions, with more degrees of surprise. It also shows that ranking functions define a qualitative ordering on events, just as possibility measures do.

Another way of interpreting ranking functions is as providing us with a way of doing order-of-magnitude probabilistic reasoning. Given a finite set W of possible worlds, choose ϵ so that ϵ is significantly smaller than 1. (The meaning of "significantly smaller" is deliberately being kept vague.) Then we can think of sets U such that $\kappa(U) = k$ as having probability of roughly ϵ^k—more precisely, of having probability $\alpha\epsilon^k$ for $\alpha > 0$ but significantly smaller than $1/\epsilon$ (so that $\alpha\epsilon^k$ is significantly smaller than ϵ^{k-1}). With this interpretation, the assumptions that $\kappa(W) = 0$ and $\kappa(U \cup U') = \min(\kappa(U), \kappa(U'))$ make perfect probabilistic sense (modulo the vagueness regarding the meaning of "significantly smaller"). We can actually make this viewpoint completely rigorous by using ideas from *nonstandard analysis*, taking ϵ to be an *infinitesimal*—a number that is closer to 0 than any real number. (In nonstandard analysis, infinitesimals are given a precise meaning. For those familiar with nonstandard analysis, we fix an infinitesimal ϵ and define $\kappa(U)$ to be the smallest natural number k such that the probability of U is at least $\alpha\epsilon^k$ for some standard real $\alpha > 0$. This is explored further in Exercise 3.24.) However, in more practical order-of-magnitude reasoning, it may make more sense to think of ϵ as a very small positive real number.

3.6 Plausibility Measures

I conclude this overview of quantitative approaches to describing likelihood by considering an approach that is a generalization of all the approaches mentioned so far, both qualitative and quantitative. This approach uses what are called *plausibility measures*, which unfortunately are distinct from the plausibility functions used in the Dempster-Shafer approach (although plausibility functions are instances of plausibility measures). I hope the reader will be able to sort through any confusion caused by this overloading of terminology.

The basic idea behind plausibility measures is straightforward. Just as with probability, we start with a set W of worlds and an algebra $\mathcal{F}$ of measurable subsets of W. A probability measure maps elements in $\mathcal{F}$ to $[0,1]$. A *plausibility measure* is more general; it maps elements in $\mathcal{F}$ to some arbitrary partially ordered set. If Pl is a plausibility measure, then we read $\mathrm{Pl}(U)$ as "the plausibility of set U". If $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$, then V is at least as plausible as U. Because the ordering is partial, it could be that the plausibility of two different sets is incomparable. An agent may not be prepared to say of two sets that one is more likely than another or that they are equal in likelihood.

Formally, a *plausibility space* is a tuple $S = (W, \mathcal{F}, \mathrm{Pl})$, where W is a set of worlds, $\mathcal{F}$ is an algebra of subsets of W, and Pl maps sets in $\mathcal{F}$ to some domain D of *plausibility values* partially ordered by a relation $\leq_D$ (so that $\leq_D$ is reflexive, transitive, and anti-symmetric) that contains two special elements $\top_D$ and $\bot_D$ such that $\bot_D \leq_D d \leq_D \top_D$ for all $d \in D$. In the case of probability measures, $\top_D$ and $\bot_D$ are 1 and 0, respectively. As usual, the ordering is defined $<_D$ by taking $d_1 <_D d_2$ if $d_1 \leq_D d_2$ and $d_1 \neq d_2$. I omit the subscript D from $\leq_D$, $<_D$, $\top_D$ and $\bot_D$ whenever it is clear from context.

There are three requirements on plausibility measures. The first two are analogues of requirements that hold for the all other notions of uncertainty that we have seen so far: the whole space gets the maximum plausibility and the empty set gets the minimum plausibility. The third requirement says that a set must be at least as plausible as any of its subsets.

Pl1. $\mathrm{Pl}(W) = \top_D$.

Pl2. $\mathrm{Pl}(\emptyset) = \bot_D$.

Pl3. If $U \subseteq V$, then $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$.

Clearly probability measures, lower and upper probabilities, inner and outer measures, Dempster-Shafer belief functions, possibility measures, and ranking functions are all instances of plausibility measures. Notice that in the case of ranking functions, the ordering is the opposite of the standard ordering on the natural numbers. That is, if κ is a ranking function, then (W, κ) is a plausibility space; $I\!N^*$, the domain of plausibility values, is ordered by $\leq_{I\!N^*}$, where $x \leq_{I\!N^*} y$ if and only if $y \leq x$ under the usual ordering on the natural numbers and ∞.

In all these cases, the plausibility ordering is total. But there are also cases of interest where the plausibility ordering is *not* total. For example, given a set $\mathcal{P}$ of probability measures on W, let D consist of all functions from $\mathcal{P}$ to $[0,1]$. The standard pointwise ordering on functions—that is, $f \leq g$ if $f(\mu) \leq g(\mu)$ for all $\mu \in \mathcal{P}$—gives a partial order on D. Given a set U, define $f_U \in D$ by taking $f_U(\mu) = \mu(U)$ for $\mu \in \mathcal{P}$; let $\mathrm{Pl}(U) = f_U$. This gives a way of associating with every set a plausibility value in D. In the case of the marbles in Example 3.2.2, *red*, *blue*, and *yellow* would have pairwise incomparable plausibilities under this approach. However, consider a variant of Example 3.2.2 where we don't know how many red, blue, or yellow marbles there are in the bag, but we do know that there more blue marbles than yellow marbles. This would correspond to $\mathcal{P}_2'' = \{\mu : \mu(\mathit{blue}) \geq \mu(\mathit{yellow})\}$. It is easy to see that

$$(\mathcal{P}_2'')_*(\mathit{blue}) = (\mathcal{P}_2'')_*(\mathit{yellow}) = 0 = (\mathcal{P}_2'')_*(\mathit{red}) = 0.$$

Nevertheless, with this plausibility measure, *blue* is more plausible than *yellow*, which seems reasonable, while *blue* and *red* are incomparable in plausibility, as are *yellow* and *red*.

The ordering on sets generated by an ordering on worlds discussed in Section 2.3 also defines a plausibility measure in the obvious way. Again, in this case, the plausibility ordering is partial; the plausibility of two sets can be incomparable. Note that plausibility measures give us a general way to describe any partial order on subsets of a space W, not just one induced by an ordering on worlds.

It should be clear that plausibility measures are very general. But what does this generality buy us? Perhaps the point is best understood in terms of epistemic knowledge. There we started with a binary relation $\mathcal{K}$ on possible worlds. We saw that by placing various restrictions on $\mathcal{K}$, we could capture various properties of knowledge. Similarly, with plausibility measures, by imposing various restrictions beyond Pl1–3 (some of which may also involve imposing various requirements on the domain on plausibility values) we can get properties of reasoning about uncertainty that are of interest to us. For example, we may be interested in plausibility measures that totally order all events, or we may want to get an analogue to P2 by assuming that there is some operation $\oplus$ on the domain of plausibility values such that $\mathrm{Pl}(U \cup V) = \mathrm{Pl}(U) \oplus \mathrm{Pl}(V)$ if U and V are disjoint. In the case of probability measures, $\oplus$ is $+$; in the case of possibility measures, it is max; in the case of ranking functions, it is min. However, there is no analogue to $\oplus$ for belief functions or lower probability measures. Thus, in the most general setting, I do not want to assume that there is necessarily such a function $\oplus$. We should only assume it if we "need" it; we should also understand the consequences of assuming it. Plausibility measures are of interest in part because they allow us to investigate such questions. I return to this topic in Sections 4.8 and 6.2.

3.7 Choosing a Representation

Before concluding this chapter, a few words are in order regarding the problem of modeling a real-world situation. It should be clear that, whichever approach is used to model uncertainty, it is important to be sensitive to the implications of using that approach. Different approaches are more appropriate for different applications. For example, as we shall see, possibility measures and ranking functions deal well with default reasoning. And in modeling decision-making behavior of agents, the non-additive behavior of belief functions may better represent how agents model uncertainty when making decisions than probability. But even before we decide what approach we should use to model a problem, there is in often another, more subtle, problem, that we have to deal with. That is the choice of possible worlds.

All the methods presented here assume that there is a set of possible worlds and some way of representing the relatively likelihood of subsets of these worlds. But how do we decide what set of worlds to consider?

Let us go back to the case of throwing a fair die. We took the set of possible worlds in that case to consist of six worlds, intuitively, one for each possible way the die might have landed. Note that there are (at least) two major assumptions being made here. The first is that all that matters is how the die lands. If, for example, we thought that the gods had a significant impact on the outcome (if the gods are in a favorable mood, then we are more likely to get a high number), then we should include the gods' mood as part of the description of a possible world. More realistically, perhaps, if we think that the die may not be fair, then we need to include its possible bias as part of the description of a possible world. There will be a possible world corresponding to each possible (bias, outcome) pair. The second assumption being made is that the only outcomes possible are 1, ..., 6. While this may seem reasonable, my experience playing games involving dice with my children in their room, which has a relatively deep pile carpet, is that a die often lands on edge.

The point of this discussion is that the choice of the set of possible worlds is often quite a nontrivial one, and encodes many of the assumptions the modeler is making about the domain. The problem gets even more difficult if we have many agents, and part of the uncertainty in the world involves what information the other agents have. The problem of choosing an appropriate set of possible worlds is not one that is typically discussed in texts on probability (or other approaches to modeling uncertainty), and it deserves more care than it is usually given. Of course, there is not necessarily a single "right" set of possible worlds to use. For example, even if the modeler thinks there is a small possibility that the coin may not be fair, it might make sense to ignore this possibility in favor of getting a simpler, but still quite useful, model of the situation. In Sections **??** and 8.1, I give some tools that may help a modeler in deciding on an appropriate set of possible worlds in a disciplined way. But even with these tools, deciding which possible worlds to consider often remains a difficult task.

While I chose the examples in this book to bring out some of the subtleties, they are still much, much simpler than many that arise in the real world. Just a word of warning ...

Exercises

3.1 Show using P2 that if $U \subseteq V$, then $\mu(U) \leq \mu(V)$.

* **3.2** Let $\alpha_U = \sup\{\beta :$ the agent prefers (U, β) to $(\overline{U}, 1-\beta)$. Show that the agent prefers (U, α) to $(\overline{U}, 1-\alpha)$ for all $\alpha < \alpha_U$ and prefers $(\overline{U}, 1-\alpha)$ to (U, α) for all $\alpha > \alpha_U$. Moreover, if U_1 and U_2 are disjoint sets, show that if the agent is rational, then $\alpha_{U_1} + \alpha_{U_2} = \alpha_{U_1 \cup U_2}$. More precisely, show that if $\alpha_{U_1} + \alpha_{U_2} \neq \alpha_{U_1 \cup U_2}$, then there is a set of bets (on U_1, U_2, and $U_1 \cup U_2$) that the agent should be willing to accept according to her stated preferences, according to which she is guaranteed to lose money. Show exactly where the assumption that if the agent prefers (U_i, α_i) to (V_i, β_i) for each $i = 1, \ldots, k$ in isolation, then the agent prefers the collection of bets $(U_1, \alpha_1), \ldots, (U_k, \alpha_k)$ to the collection $(V_1, \beta_1), \ldots, (V_k, \beta_k)$ comes into play.

3.3 Show that if W is finite then $\mu_*(U) = \mu(V_1)$, where $V_1 = \cup_{B \subseteq U,\, B \in \mathcal{F}} B$ and $\mu^*(U) = \mu(V_2)$, where $V_2 = \cap_{U \subseteq B,\, B \in \mathcal{F}} B$.

* **3.4** Prove Theorem 3.2.3.

3.5 Show that inner and outer measures satisfy Equations (3.1) and (3.2).

3.6 Prove the inclusion-exclusion rule (Equation 3.5). (Hint: use induction on n, the number of sets in the union.)

* **3.7** Show that if μ is a σ-additive probability measure on a σ-algebra $\mathcal{F}$, then for every set U (not necessarily in $\mathcal{F}$), there exists a set $U_{\mathcal{F}} \in \mathcal{F}$ such that $U_{\mathcal{F}} \subseteq U$ and $\mu(U_{\mathcal{F}}) = \mu_*(U)$. Moreover, show that, without loss of generality, we can assume that $(U \cap U')_{\mathcal{F}} = U_{\mathcal{F}} \cap U'_{\mathcal{F}}$. (If W is finite, this result follows easily from Exercise 3.3. Indeed, that exercise shows that $U_{\mathcal{F}}$ can be taken to be $\cup_{U \subseteq B,\, B \in \mathcal{F}} B$. This result holds even if W is infinite, but we need the assumption that $\mathcal{F}$ is a σ-algebra and μ is countably additive. Note that if U is infinite, $\cup_{U \subseteq B,\, B \in \mathcal{F}} B$ is not necessarily in $\mathcal{F}$, even if $\mathcal{F}$ is a σ-algebra, since the union may be over uncountably many sets.)

3.8 Prove Equation (3.6) for inner measures. (You may assume the results of Exercises 3.6 and 3.7.)

3.9 Show that lower and upper probabilities satisfy analogues of Equations (3.1) and (3.2) (with μ_* and μ^* replaced by $\mathcal{P}_*$ and $\mathcal{P}^*$, respectively), but show by means of a counterexample that they do not satisfy the analogue of Equation (3.6) in general. (Hint: it suffices for the counterexample to consider three possible worlds and a set consisting of two probability measures.)

3.10 Let $W = \{w, w'\}$ and define $\mathrm{Bel}(\{w\}) = 1/2, \mathrm{Bel}(\{w'\}) = 0, \mathrm{Bel}(W) = 1$, and $\mathrm{Bel}(\emptyset) = 0$. Show that Bel is a belief function, but there is no probability μ that we can place on W such that $\mathrm{Bel} = \mu_*$.

3.11 Construct two belief functions Bel_1 and Bel_2 on $\{1, 2, 3\}$ such $\mathrm{Bel}_1(i) = \mathrm{Bel}_2(i) = 0$ for $i = 1, 2, 3$ (so that Bel_1 and Bel_2 agree on singleton sets) but $\mathrm{Bel}_1(\{1, 2\}) \neq \mathrm{Bel}_2(\{1, 2\})$.

3.12 Show that $\mathrm{Bel}(U) \leq \mathrm{Plaus}(U)$ for all sets U.

* **3.13** Complete the proof of Theorem 3.3.1.

* **3.14** Prove Theorem 3.3.2. (Hint: proving that Bel_m is a belief function requires proving B1, B2, and B3. B1 and B2 are obvious, given M1 and M2. For B3, proceed by induction on n, the number of sets in the union, using the fact that $\mathrm{Bel}_m(A_1 \cup \ldots \cup A_{n+1}) = \mathrm{Bel}_m((A_1 \cup \ldots \cup A_n) \cup A_{n+1})$. To construct m given Bel, define $m(\{w_1, \ldots, w_n\})$ by induction on n so that Equation (3.7) holds.)

3.15 Suppose that m_1 and m_2 are mass functions, Bel_1 and Bel_2 are the corresponding belief functions, and there do not exist sets U_1 and U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. Show that there must then be sets V_1, V_2 such that $\mathrm{Bel}_1(V_1) = \mathrm{Bel}_2(V_2) = 1$ and $V_1 \cap V_2 = \emptyset$.

3.16 Show that $\oplus$ is commutative and associative, and that m_{vac} is the neutral element for $\oplus$.

3.17 Poss3 says that $\mathrm{Poss}(U \cup V) = \max(\mathrm{Poss}(U), \mathrm{Poss}(V))$ for U, V disjoint. Show that $\mathrm{Poss}(U \cup V) = \max(\mathrm{Poss}(U), \mathrm{Poss}(V))$ even if U and V are not disjoint. Similarly, if κ is a ranking function, show that $\kappa(U \cup V) = \min(\kappa(U), \kappa(V))$ even if U and V are not disjoint.

3.18 Show that $\mathrm{Nec}(U) \leq \mathrm{Poss}(U)$ for all sets U.

3.19 Prove Theorem 3.4.1.

3.20 Show that $\max(\mu(U), \mu(V)) \leq \mu(U \cup V) \leq \min(\mu(U) + \mu(V), 1)$. Moreover, show that these bounds are optimal, in that there is a probability measure μ and sets U_1, V_1, U_2, and V_2 such that $\mu(U_1 \cup V_1) = \max(U_1, V_1)$ and $\mu(U_2 \cup V_2) = \min(\mu(U) + \mu(V), 1)$.

3.21 Suppose that Poss is a possibility measure on W and define a partial order $\succeq$ on W such that $w \succeq w'$ if $\mathrm{Poss}(w) \geq \mathrm{Poss}(w')$.

(a) Show that $U \succeq^s V$ implies $\mathrm{Poss}(U) \geq \mathrm{Poss}(V)$.

(b) Show that $\mathrm{Poss}(U) > \mathrm{Poss}(V)$ implies $U \succ^s V$.

(c) Show that the converses to parts (a) and (b) do not hold in general. (Hint: consider the case where one of U or V is the empty set.)

(d) Show that if $\mathrm{Poss}(w) > 0$ for all $w \in W$, then the converses to parts (a) and (b) do hold.

Analogous results can be proved for ranking functions, using almost identical proof techniques.

3.22 Show directly (without using Proposition 2.3.4) that Poss is qualitative; that is, if $\text{Poss}(U_1 \cup U_2) > \text{Poss}(U_3)$ and $\text{Poss}(U_1 \cup U_3) > \text{Poss}(U_2)$, then $\text{Poss}(U_1) > \text{Poss}(U_2 \cup U_3)$. An almost identical argument works to show that ranking functions are qualitative.

3.23 Show that Poss_κ (as defined in Section 3.4) is a possibility measure.

3.24 Suppose that we add an "infinitesimal" element ϵ to the real numbers. Informally, think of ϵ as a very small number. Multiplication of small numbers results in even smaller numbers (for example, $.01 \times .01 = .0001$). Thus, we can think of ϵ^2 as much smaller than ϵ and, in general, $\epsilon^{k'}$ as being much than ϵ^k if $k < k'$. "Much smaller" here means that $a\epsilon^k > \epsilon^{k'}$ for all real numbers $a > 0$ if $k' > k$. Elements of these "extended reals" are polynomials in ϵ, and thus have the form $a_0 + a_1\epsilon + a_2\epsilon^2 + \cdots + a_k\epsilon^k$, where $a_0, \ldots, a_k$ are real numbers. We add, subtract, and multiply polynomials as usual. A nonstandard probability measure on W can now be viewed as a mapping μ associating with each measurable subset of W an extended real, satisfying P1 and P2. In this framework, if $\mu(U) = a_0 + a_1\epsilon + a_2\epsilon^2 + \cdots + a_k$, then we can define $\kappa(U)$ to be the least k such that $a_k \neq 0$. (If $\mu(U) = 0$, then $\kappa(U) = \infty$.) Note that if $\kappa(U) = k < \infty$, then $a_k > 0$, by our assumption that $\epsilon^k \gg \epsilon^{k'}$ if $k < k'$. This definition of κ trivially satisfies R1 and R2. Show that it also satisfies R3.

Notes

There are many, many texts on all facets of probability; three standard introductions are by Ash [1970], Feller [1957], and Halmos [1950].

The use of the principle of indifference in probability is associated with a number of people in the 17th and 18th century, chief among them perhaps being Bernoulli and Laplace. See Hacking [1975] for a historical discussion. The term *principle of indifference* is due to Keynes [1921]; it has also been called the *principle of insufficient reason* [Kries 1886].

Many justifications for probability can be found in the literature. As was stated in the text, the strongest proponent of the relative-frequency interpretation was von Mises [1957]. A recent defense of this position was given by van Lambalgen [1987].

Ramsey's [1931] is perhaps the first careful justification of the subjective viewpoint; the variant of his argument given here is due to Paris [1994]. De Finetti [1931, 1937, 1972] proved the first Dutch Book arguments.

Another famous justification of probability is due to Cox [1946], who showed that any function that assigned degrees to events that satisfied certain minimal properties (such as the degree of belief in $\overline{U}$ is a decreasing function in the degree of belief in U) must be isomorphic to a probability measure. Unfortunately, Cox's argument is not quite correct as stated; his hypotheses need to be strengthened (in ways that make them less compelling) to make it correct [Halpern 1999a; Halpern 1999b; Paris 1994].

Yet another justification for probability is due to Savage [1954], who showed that a rational agent (where "rational" is defined in terms of a collection of axioms) can, in a precise sense, be viewed as acting as if his beliefs were characterized by a probability measure. More precisely, Savage showed that a rational agent's preferences on a set of actions can be represented by a probability measure on a set of possible worlds combined with a utility function on the outcomes of the actions; the agent then prefers action a to action b if and only if the expected utility of a is higher than that of b. (See Section **??**.) Savage's approach has had a profound impact on the field of *decision theory* [Kreps 1988].

The idea of modeling imprecision in terms of sets of probability functions is an old one, apparently going back as far as Boole [1854, Chapters 16–21]. Borel [1943, Section 3.8] suggested that upper and lower probabilities could be measured behaviorally, as betting rates on or against an event. These arguments were formalized by Smith [1961]. Walley [1991] provides a thorough discussion of the use of what he calls *upper* and *lower previsions*. These are upper and lower bounds on the uncertainty of an event, but are not defined in terms of sets of probability measures. The idea of using inner measures to capture imprecision is due to Fagin and me [1991b]. The inclusion-exclusion theorem is discussed in most standard probability texts, as well as in standard introductions to discrete mathematics (for example, [Maurer and Ralston 1991]).

Belief functions were originally introduced by Dempster [1967, 1968], and then extensively developed by Shafer [1976]. Choquet [1953] independently introduced the notion of *capacities* (now often called *Choquet capacities*); infinitely-monotone capacities are mathematically equivalent to belief functions. Theorem 3.3.1 was originally proved by Dempster [1967], while Theorem 3.3.2 was proved by Shafer [1976, p. 39]. Examples 3.3.3 and 3.3.4 are taken from Gordon and Shortliffe [1984] (with slight modifications). Fagin and I [1991b] and Ruspini [1987] were the first to observe the connection between belief functions and inner measures. Exercise 3.8 is Proposition 3.1 in [Fagin and Halpern 1991b]; it also follows from a more general result proved by Shafer [1979]. Shafer [1990] discusses various justifications for and interpretations of belief functions. He explicitly rejects the idea of belief function as a lower probability.

The idea of a possibility measure was introduced by Zadeh [1978], who developed the idea from his earlier work on fuzzy sets and fuzzy logic [Zadeh 1975]. The theory was greatly developed by Dubois, Prade, and others; a good introduction can be found in [Dubois and Prade 1990]. Theorem 3.4.1 on the connection between possibility measures and consonant plausibility functions is proved, for example, by Dubois and Prade [1982].

Ordinal conditional functions were originally defined by Spohn [1988], who allowed them to have values in the ordinals (the *ordinal* numbers go beyond the natural numbers, and allow us to talk about different types of infinity), not just values in $\mathbb{N}^*$. Spohn also showed the relationship between his ranking functions and nonstandard probability, as sketched in Exercise 3.24. (For more on nonstandard probability, see [Davis 1977].) The degree-of-surprise interpretation for ranking functions goes back to Shackle [1969].

Friedman and I [1995, 1998] introduced plausibility measures; the discussion in Section 3.6 is taken from our papers. Weber [1991] independently introduced an equivalent notion.