Chapter 1

Fundamentals of Probability and Judgement

1.1 Introduction

This book concerns the elicitation of expert knowledge in probabilistic form. Before we can discuss what this means and the techniques for doing it, we need to explore some fundamental facts about probability and the way in which people formulate judgements of probability. This chapter begins with an introduction to probability and elicitation. It continues with a discussion of the nature of probability, arguing that, for the kinds of uncertain quantities for which expert opinion is typically sought, the usual understanding of probability in terms of long-run repetition of events is inadequate. We then consider how experts construct probability judgements, and find that probabilities are not pre-formed numbers just waiting to be expressed. On the contrary, psychological research tells us that judgements are formed 'on the fly' in response to questioning about uncertain quantities and are likely to be highly context dependent. Finally, we ask how such probability judgements might relate to the normative theories that underpin the interpretation and use of probabilities in statistics, decision theory and risk analysis.

1.2 Probability and elicitation

1.2.1 Probability

The probability of an event is a measure of how likely it is to occur. Probability 0 means that the event is certain not to occur, whereas probability 1 means that it is

Uncertain Judgements – Eliciting Experts' Probabilities A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow

certain to occur. Values from 0 to 1 describe increasing chances that the event will occur. The central value, 0.5, represents an event that is as likely to occur as it is not to occur. Events with probabilities above 0.5 are more likely to occur than not to occur, and conversely events with probabilities below 0.5 are more likely not to occur than to occur.

The symbol that is almost universally adopted to denote probability is P. Thus, if E is an event, then $P(E)$ denotes the probability of that event. For example, if E is the event of getting the result 'Heads' in a toss of an ordinary coin, then we can say $P(E) = 0.5$, because it is generally agreed in this situation that 'Heads' and 'Tails' are equally likely to occur. Similarly in the roll of a die ('die' here is the singular of 'dice'), there are 6 equally likely results and if S is the event of getting a 6 then $P(S) = \frac{1}{6}$.

The theoretical study of probability is a branch of mathematics that deals with laws and theorems about how probabilities behave and combine. For instance, suppose that events E and F are mutually exclusive. The term 'mutually exclusive' is defined in the Glossary; it simply means that E and F both cannot occur. If one occurs, then the other cannot. Let $EorF$ be the event that either E or F occurs. Then one of the fundamental laws of probability theory (the Addition Law) is that $P(EorF) = P(E) + P(F)$.

The statement that E and F are mutually exclusive implies that if I know that E has occurred, then the probability that F occurs must be zero. The occurrence of E changes the probability of F. For the inexperienced observer, one of the most difficult aspects of probability (and the source of some perplexing paradoxes) is the manner in which the probability of an event is affected by other events or other information that we might have. In this case, we need to distinguish between the probability of F when we do not know whether E has occurred and its probability when we do have that information. The first is just $P(F)$, and is called the *unconditional* probability of F. But if we know that E has occurred we have $P(F \mid E) = 0$, and this is the *conditional* probability of F *given* E. Another example of conditional probability can be found in the toss of the die. If E denotes the event that the result is an even number, then $P(S \mid E) = \frac{1}{3}$; that is, given that the result is an even number (2, 4 or 6) the probability of getting a 6 is one-third.

A more complex example is the probability that a specified person is killed in a road accident in the next 12 months. If we know nothing about the person except that he/she lives in England, then we could assess that probability as about one in 20,000 (because, although figures are not readily available for England alone, about 3000 people are killed on British roads each year). If, however, we know that the person is aged between 17 and 21, then the probability is larger, because this age group has more accidents. If we also know that the person is male, the probability increases again. A person's chance of being killed on the road varies with their age and gender, where they live in England, their occupation, whether they are married, and so on.

Pursuing this example further, what is the probability that I will be killed in a road accident in the next 12 months? If we consider all the relevant conditioning

factors – my age, gender, location, marital status, the model of car that I drive, the number of miles that I drive each year, and so on – then there is nobody else in England (and never has been) with exactly the same characteristics. There will therefore be no data on which to assess that probability, and it is even questionable how to define it. We will explore these issues more thoroughly in Sections 1.3.2 and 1.3.3, but it is already clear why probabilities can be confusing for ordinary people. One reason why road safety advice (such as to wear seat belts, not to use mobile phones while driving or to drive more slowly in conditions of poor visibility) often has limited effect is because people do not see it as necessarily applying to them personally. They can believe that using mobile phones is dangerous in general, but that they personally can do so safely. It may not be rational to believe that they are special in this way, but it is certainly true that each person's individual characteristics will condition the probability and make them more or less at risk than the average person.

1.2.2 Random variables and probability distributions

Uncertainty about a single event is quantified by its probability. We are often interested, however, not so much in an uncertain event as in an uncertain quantity. An uncertain quantity is usually called a *random variable*. An event either occurs or does not occur, whereas a random variable may take any value of some collection of possible values. If the variable may take any value within some range, then we call it a *continuous* random variable. An example is the weight of the Great Pyramid at Giza, Egypt, which could, in principle, be any positive value (although clearly it would be very many tons). In contrast, a *discrete* random variable can only take certain distinct values, and cannot have values between these. An example is the number of stones in the Great Pyramid, which, in principle, could be 0, 1, 2 or any other positive integer value (although again it is clearly a large number), but not, for instance, any value between 0 and 1 or between 2,000,000 and 2,000,001.

Uncertainty about a random variable X is described by specifying the probability $P(X \leq x)$ for any x. So, although we can no longer characterise uncertainty about a random variable with a single probability, the description is still in terms of probabilities. (Note that $X \leq x$ is an event, the event that the true value of X is less than or equal to x.) If we think of $P(X \leq x)$ as a function of x, then it is called the (cumulative) *distribution function* of X. Examples of these functions for both discrete and continuous random variables are shown in the Glossary. Note that we can also have conditional distributions. For instance, the conditional distribution of X given some event E is specified by the probabilities $P(X \leq x \mid E)$ for any x.

While the distribution function is formally the way to define the probability distribution of a random variable, there are alternative formulations that are more intuitive, but which differ for continuous and discrete random variables. For a discrete random variable, it is more natural to use the set of probabilities $P(X = x)$ that give the probability that X will take each of its possible values. This is called the *probability mass function* of X. Figure 1.1 shows the probability mass functions

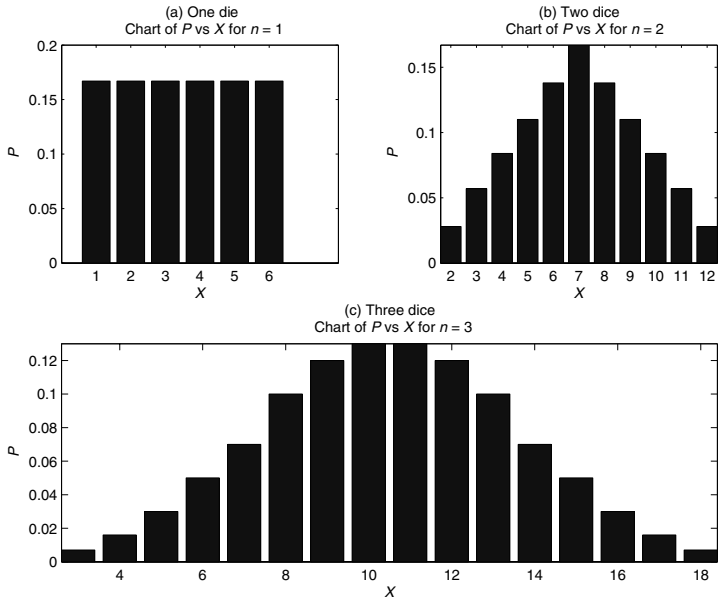

Figure 1.1: Probability mass functions for total score on n dice.

for three random variables. In Figure 1.1(a), we see that of the score on the toss of a single die. Figure 1.1(b) shows the probability mass function for the total score on tossing two dice and Figure 1.1(c), the same for the total of three dice.

The score on one die is equally likely to be 1, 2, 3, 4, 5 or 6, and this appears in Figure 1.1(a) as a distribution of uniform height. The same for two dice in Figure 1.1(b) has a triangular shape, while that of three dice in Figure 1.1(c) climbs, flattens out and then falls smoothly.

For continuous random variables, a similar picture is shown by the *probability density function* (pdf). Figure 1.2 shows some typical pdfs.

Both density functions are *unimodal*, meaning that they rise to a single peak (mode) before falling again. The density in Figure 1.2(a) is *symmetric* (like those in Figure 1.1), while that in Figure 1.2(b) is *skewed*. It should be noted that, whereas the heights of the bars in the probability mass function plots in Figure 1.1 are actual probabilities, the heights of the pdf curves in Figure 1.2 are not probabilities. Instead, it is the area under the curve between any two points, say x_1 and x_2, that is a probability; specifically, this area is $P(x_1 \leq X \leq x_2)$, the probability that X lies between x_1 and x_2.

The distributions in Figure 1.2 are examples of the many families of distributions that are used in statistics. Figure 1.2(b) is an example of a *beta* distribution; specifically it is the beta distribution with parameters 2 and 4. Figure 1.2(a) is a *normal* distribution; specifically it is the normal distribution with parameters 0

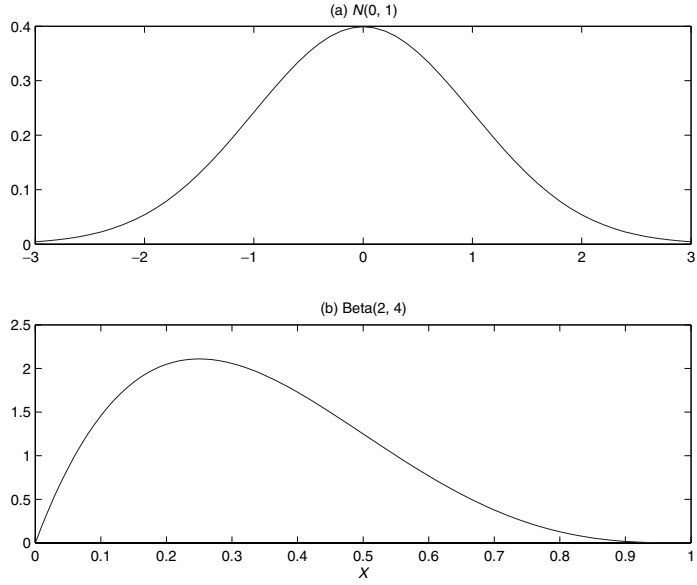

Figure 1.2: Two probability density functions.

and 1, also known as the *standard normal distribution*. Several other widely used distribution families are described in the Glossary.

1.2.3 Summaries of distributions

Although a probability distribution is defined by its distribution function, or equivalently by its probability mass function or probability density function, it is useful to have other kinds of descriptions that capture particular features of a distribution. So there are many different *summaries* that are used in statistics to present information about a distribution. Since these are also widely used in elicitation, the most important ones are listed here.

- *Probabilities*. Individual probabilities, such as $P(X = 2)$, $P(1 \leq X \leq 2)$ or $P(X \leq 10)$, are often used as summaries in their own right.

- *Quantiles*. The qth *quantile* of the distribution of X is the value x_q such that $P(X \leq x_q) = q$. The most widely used quantiles are the percentiles, median and quartiles. The nth percentile is $x_{0.01n}$. The 50th percentile, $x_{0.5}$, is known as the *median*, and it divides the range of X into two equally probable ranges (with probabilities 0.5); X is equally likely to lie above $x_{0.5}$ or below $x_{0.5}$. The lower *quartile* is the 25th percentile and the upper quartile is the 75th percentile. Together, the quartiles and the median divide the range of X

into four equiprobable regions (with probabilities 0.25). Of course, all the percentiles together divide the range into 100 equiprobable regions (all with probability 0.01). We sometimes refer to the *tertiles*, which divide the range into three equiprobable regions.

- *Intervals*. For any $s > t$, there is a probability $s - t$ that X lies in the interval (i.e., range of values) from x_t to x_s, written as $[x_t, x_s]$. It is often referred to as a $100(s - t)\%$ probability interval (or *credible interval*) for X. An interval like this with a suitably high probability, such as 90 or 95%, provides a range of values in which the true value of X will 'probably' lie.

- *Location measures*. These measures try to represent in some sense a typical, or representative, value of X. The median is a location measure, being the central value (such that X is equally likely to be higher or lower). Another measure is the *mode*, defined to be the value of X at which the probability mass function or pdf reaches its maximum. In the case of a discrete random variable, the mode is the *most probable* value for X. For a continuous random variable, this interpretation suffers from some technical ambiguities, but is still the usual way to explain the mode. The location measure that is most often used in statistical analysis is the *mean*. The mean, or expected value, of X is interpreted as the average value, or more formally, if we were able to observe the values of many random variables all with the same distribution as X, then the average of these values would be the mean. The mean has its own notation, $E(X)$ (standing for the *expectation* of X).

- *Measures of scale or dispersion*. These measures represent in different ways how far from its mean (or some other location measure) the random variable X might be. They can be seen as descriptions of how much uncertainty there is concerning X, since a large value of any of these measures implies that X may be far from any typical or representative value. The simplest is the inter-quartile range, which is the difference between the upper and lower quartiles. The most widely used measure in statistical analysis is the *variance*, which is the expected squared distance of X from its mean. Formally, we write this as $E\{(X - E(X))^2\}$. The square root of the variance is known as the *standard deviation* and is often more useful as a measure of dispersion because it is on the same scale as X.

- *Measures of shape*. Qualitative measures of shape include describing the density as unimodal, bimodal (rising to two distinct maxima, with a dip between) or multi-modal (having three or more maxima). We can also say that it is symmetric (as in Figure 1.2(a)), skewed to the right (as in Figure 1.2(b)) or skewed to the left (as in the mirror image of Figure 1.2(b)). However, there are also quantitative measures of skewness (where a symmetric distribution has value 0, a distribution skewed to the right has a positive value and a distribution skewed to the left has a negative value), and kurtosis (the tendency for the mode to be more or less sharply curved).

In order to describe a distribution effectively, a statistician will often use several summaries.

The reader may have encountered many of these summaries in a different guise, as summaries of a set of data. For instance, the mean of a sample is the average of all the values. There is a natural correspondence between the summaries of samples and the summaries of distributions, and a sample can often usefully be thought of as an *empirical* distribution.

1.2.4 Joint distributions

If we have two random variables, say X and Y, the uncertainty about them is not completely described by giving their separate distributions. The distribution of X gives us the value of $P(X \leq x)$ and that of Y gives us the value of $P(Y \leq y)$, but this is not enough to determine the *joint* probability $P(X \leq x, Y \leq y)$, which is the probability that both events, $X \leq x$ and $Y \leq y$, occur. The reason is that the occurrence of one event may change the probability of the other, in the way considered in Section 1.2.1.

Another of the laws of probability theory (the Multiplication Law) is that for two events E and F the joint probability $P(E, F)$ equals the product of the unconditional probability $P(E)$ and the conditional probability $P(F \mid E)$, that is, $P(E, F) = P(E)P(F \mid E)$. Equivalently, $P(E, F) = P(F)P(E \mid F)$. If the occurrence or non-occurrence of F does not change the probability of E then $P(E \mid F) = P(E)$ and we have $P(E, F) = P(E)P(F)$. In this situation, we say that E and F are *independent*. Notice now that this also implies that $P(F \mid E) = P(F)$: independence is a symmetric relationship, and if the occurrence or non-occurrence of F does not change the probability of E then the occurrence or non-occurrence of E does not change the probability of F.

The same ideas apply to probability distributions. If X and Y are two discrete random variables, then their joint probability mass function comprises the probabilities $P(X = x, Y = y)$ for all possible values x of X and y of Y. They are said to be independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y. Two continuous random variables are said to be independent if their joint pdf is the product of their separate pdfs. If random variables are independent then knowing their separate probability distributions *is* enough to know all about their joint uncertainty. But otherwise we need to consider that the occurrence of some particular value of X may influence the distribution of Y or, conversely, that the occurrence of any particular value of Y will influence the distribution of X.

If X and Y are not independent, we will need to consider the *conditional* distributions of X given $Y = y$ (for all possible values y) and/or the conditional distributions of Y given $X = x$ (for all possible x). Therefore, the joint uncertainty of two (or more) random variables is potentially a complex thing, and may require new kinds of summaries to describe it.

- *Measures of correlation*. These measures describe the degree to which the value of one variable influences the value of another. They take the value 0

when random variables are independent and ± 1 when they are totally dependent, meaning that as soon as we know the value of one variable there will be no uncertainty about the value of the other. In the case of total dependence, the sign of the correlation coefficient indicates which of two forms of total dependence applies. Correlation of $+1$ means that as X increases so does Y, whereas if the correlation is -1 then as X increases, Y decreases. Values between these extremes indicate greater or lesser degrees of dependence, with a positive sign indicating that higher values of one tend to be associated with higher values of the other (and negative sign meaning that higher values of one tend to be associated with lower values of the other). The usual correlation coefficient is formally known as the *Pearson correlation coefficient*. It only takes the value ± 1 if the variables are totally dependent in a linear relation (increasing X by one unit always causes Y to increase by the same amount, regardless of the original value of X). Other correlation coefficients exist that measure *rank* correlation, and give values ± 1 whenever each variable is totally dependent on the other.

Also, just as individual probabilities are used as summaries for a single variable, we may use joint or conditional probabilities to summarise the features of a joint distribution.

1.2.5 Bayes' Theorem

An important consequence of the asymmetry in the Multiplication Law of probabilities is Bayes' Theorem (named after an eighteenth-century mathematician and clergyman called Thomas Bayes). In its simplest form it states that

$$P(E \mid F) = \frac{P(E)P(F \mid E)}{P(F)}.$$

The reason this is an important result is that it provides a recipe for learning from experience. In this context, we interpret E as an uncertain event of interest and F as a piece of new information that we obtain (we learn that the event F occurs). Then Bayes' Theorem explains how to convert from the *prior probability* of E, which is $P(E)$, to the *posterior probability* $P(E \mid F)$. The words 'prior' and 'posterior' here refer to the state of knowledge before and after learning that F occurs. The conversion consists of multiplying by $P(F \mid E)/P(F)$.

What is not apparent from this simple description, but would take too much space here to explain more fully, is how this 'recipe for learning' can really be applied in practice. However, this simple result underpins a philosophy of statistical inference known as the Bayesian approach, which is characterised by using the data and a form of Bayes' Theorem to update an initial state of knowledge (the prior distribution) to a new state of knowledge (the posterior distribution).

1.2.6 Elicitation

This book concerns the elicitation of experts' knowledge about one or more uncertain quantities in probabilistic form, and we are now in a position to appreciate what this 'probabilistic form' is. It is a (joint) probability distribution for the random variable(s) in question. The purpose of such elicitation is to construct a probability distribution that properly represents the expert's knowledge/uncertainty. The person whose knowledge is to be elicited is usually referred to as an 'expert', and while in principle there is no particular reason for them to have special knowledge or expertise, the fact that someone deems it worthwhile to carry out the elicitation implies that the expert's knowledge and judgements are worth having.

Elicitation is an important activity in a variety of fields. It has been widely practised in the design and management of large, complex engineering projects. Such projects are often essentially unique, so that there is very limited experience about the performance of components individually and in combinations. It is natural then to draw on expert judgements. In particular, there has been extensive use of elicitation in connection with nuclear installations.

Similarly, elicitation has played an important role in complex decision-making. The most difficult decisions are those where the consequences are subject to substantial uncertainty, and where those uncertainties are themselves not easy to judge. The use of expert elicitation to quantify the uncertainty in key variables then feeds directly into the decision itself.

Two statistical contexts also call for elicitation. One is the design of experiments. The purpose of experiments is to gain information regarding variables about which there is substantial uncertainty. Paradoxically, however, it is important to be able to use what knowledge one has about those variables in order to plan efficient experiments.

The other statistical context is in the Bayesian approach to statistics, a vital component of which (as is suggested in Section 1.2.5) is the use of prior information to augment the information from the statistical data. See, for example, O'Hagan and Forster (2004, Chapter 6). Elicitation of prior information is accepted as having a fundamental role in Bayesian statistics. In other areas in which elicitation is practised, the expert's knowledge feeds directly into the analysis of the underlying problem and will typically influence the outcome of that analysis strongly. In Bayesian statistics, however, it will often be the case that the statistical data will contain far more information than the prior knowledge, so the prior information may not be influential. Formal elicitation of prior distributions in Bayesian statistics has been used only in situations where prior information is appreciable and the data limited.

Numerous examples of all these contexts for elicitation will be found in Chapter 10.

1.3 Uncertainty and the interpretation of probability

1.3.1 Aleatory and epistemic uncertainty

The essence of elicitation is to capture an expert's knowledge about some uncertain quantity in a probability distribution that appropriately recognises the degree of uncertainty. It is useful to identify two different kinds of uncertainty that are sometimes known by the terms *aleatory* and *epistemic* uncertainty.

Aleatory uncertainty is induced by randomness. The word 'aleatory' derives from the Latin *alea*, meaning a die (singular of 'dice', readers may know the Latin quotation *alea jacta est* – the die is cast – attributed to Julius Caesar on crossing the Rubicon). Wherever we are interested in characterising uncertainty in one or more instances of a random process, then aleatory uncertainty is present. Epistemic uncertainty is due to imperfect knowledge about something that is not in itself random and is, in principle, knowable. The word 'epistemic' is Greek and means 'pertaining to knowledge'.

Consider, for example, the improvement in lung function that might be produced by a drug for asthma sufferers. The most widely used measure of lung function is FEV_1, which is the amount of air that the patient can expel in one second with maximum effort. If we ask an expert to assess the FEV_1 value that an individual patient will achieve using the drug, then she will have uncertainty about this value for a variety of reasons. (Note that we are adopting a convention here, which is explained in Section 2.3, that the expert is female.) First, there is aleatory uncertainty due to the fact that an individual patient will produce different FEV_1 readings when given repeat lung function tests. This is unavoidable random variation. Second, if we suppose that the expert is being asked about an unspecified, randomly chosen individual, then there is also aleatory uncertainty due to variability between patients. In addition, there is epistemic uncertainty because of various things that the expert has imperfect knowledge of. These may include uncertainty about how much within-patient variability there is in repeated FEV_1 measurements or uncertainty about how FEV_1 varies between patients. Even if the expert has enormous experience of both between- and within-patient variability of FEV_1 readings, she is likely to be uncertain about the extent of improvement that is achieved by the drug.

Statisticians usually separate the two kinds of uncertainty in the statistical models that they build. For the above example, we could characterise the single FEV_1 reading y as

$$y = \mu + \alpha + \tau p + \sigma e,$$

where μ is the mean level of FEV_1 for untreated patients, α is the effect of the drug in terms of the mean increase in FEV_1 that it produces, τ is the standard

deviation of between-patient variability, σ is the standard deviation of within-patient (i.e., between measurements) variability, and p and e are zero-mean, unit-variance random variables that we might assume to be normally distributed. In this expression, it is p and e that represent the aleatory uncertainties. These give y a random addition (positive or negative) for the individual patient and the individual measurement. The other symbols, μ, α, τ and σ, represent epistemic uncertainties. Unless the expert has considerable practical experience, they will all be uncertain, but, in principle (given enough data), they are knowable. Statisticians refer to μ, α, τ and σ as *parameters*. A statistical model can be viewed as a representation of data in terms of (aleatory) probability distributions and (epistemic) parameters.

1.3.2 Frequency and personal probabilities

The distinction between aleatory and epistemic uncertainties is paralleled by the distinction between two different definitions of probability. *Frequency* probability is the definition that almost all people learn when they first encounter theories of probability and statistics. According to the frequency definition, the probability of an event is the proportion of times that it occurs if we conduct a long sequence of repetitions. Thus, the probability of obtaining 6 on a single toss of a die is defined to be the proportion of times that 6 would occur if we tossed it an infinite number of times. This definition is essentially only applicable to aleatory uncertainties, because it requires events to be repeatable in a process having intrinsic randomness. This is obviously true of tossing a die and is also true of making repeated measurements of FEV_1 on an individual patient or FEV_1 measurements on a series of randomly chosen patients. The definition cannot, however, apply to the effect of the drug. We cannot imagine this to be repeatable, since this is a specific drug and would not be completely equivalent to any other.

Epistemic uncertainties are typically associated with one-off, unrepeatable things. The same is almost always true of parameters in statistical models. If we wish to express epistemic uncertainty through probabilities, we must find another definition.

The answer is to use *personal* probability, also sometimes called subjective probability. According to this definition, probability represents someone's *degree of belief* in an uncertain proposition. This applies to both aleatory and epistemic uncertainties. I have, for instance, a degree of belief in whether a toss of a die will yield a 6, and I can have a degree of belief in the proposition that a particular drug will increase FEV_1 for asthma patients on average by 100 ml or more.

It is clear that the terms 'personal' or 'subjective' are appropriate because my degree of belief in one of these propositions may be different from yours. This may not be true for something as simple as a toss of a die but for epistemic uncertainties (which are associated with imperfect knowledge) probabilities will always depend on what knowledge a person has. In everyday usage, the word 'subjective' has

unfortunate connotations of opinions contaminated by personal bias, prejudice and even irrationality or superstition. It is important to recognise that the objective of good elicitation is to eradicate such elements and to structure the process of elicitation in such a way as to assist the expert in rational and thoughtful evaluation of her knowledge and experience. The expert inevitably has different knowledge from others, so her probabilities are personal, but they should not be 'subjective' in any of those pejorative senses.

1.3.3 An extended example

To help clarify the distinctions and ideas in Sections 1.3.1 and 1.3.2, it is useful to consider another example in some detail.

Suppose that a timber company is considering planting a species of tree that it has not previously used. It asks an expert for her judgement of what yield it will get if it plants this species. (For the purposes of this example, we will define the yield to be the volume of usable timber per tree, although in forestry the more usual definition is volume per hectare per year.) An important first distinction is between the yield of a single tree and the average over all trees that the company might plant. This is known in statistics as the distinction between an *individual* sampled observation and the underlying *population* mean. In this case, the population is the collection of all the trees that the company will grow if it decides to use this species, and an individual tree will usually be regarded as being randomly drawn from this population. The yield of an individual tree therefore has aleatory uncertainty that is described by the *distribution* of yields in the population. For instance, if 30% of trees in the population yield more than $50\,m^3$ of timber, then there is a probability of 0.3 that an individual tree will yield more than $50\,m^3$. The aleatory uncertainty is completely described if we know this distribution of yields in the population.

However, there is another source of uncertainty – that this distribution is not known. The yields of trees of this species will have been observed in other places where it has been grown, and this is likely to form a part of the expert's knowledge, but there is uncertainty about how well the species will grow on this company's land. Furthermore, it is this distribution, and particularly the mean of the distribution, that is of interest to the company. It is the mean yield, the average over all the trees, that relates directly to the profitability of this species, and to the decision whether to plant it. It is this mean yield that the expert is asked to assess, not yields of individual trees.

Is the expert's uncertainty about mean yield aleatory or epistemic? We could think of the company's land as just one of the many sites where this species has been and might be grown. There is then another level at which we can conceive a population, the population of sites, and a distribution of mean yields over these sites. So if 25% of sites have mean yields of more than $45\,m^3$ per tree, then we might suggest that the probability of the company's site producing a mean yield over $45\,m^3$ per tree is 0.25. This presents the uncertainty about mean yield as aleatory, but such an interpretation is *not* appropriate. The company's site cannot

be regarded as randomly drawn from the population of sites. We know its latitude, its altitude, the nature of the geology and topography, all of which make this site different from others. It is factors such as these that the expert will be expected to take into account, in addition to any knowledge about the yield of this species at other sites.

The uncertainty about mean yield in this site is predominantly epistemic because it is *not* a randomly chosen site. The uncertainty derives from lack of knowledge about how the specific features of this site will affect the mean yield. Whereas the frequency interpretation of probability is adequate to describe the distribution of yields in a population of trees, it cannot apply to the mean yield of a specific site, because that site is a 'one-off'. There is no other site that is exactly like it, and when we use probability to describe the expert's uncertainty about the mean yield, the only meaningful interpretation for those probabilities is the personal or subjective interpretation.

Note that if the expert were to be asked about the yield of an individual tree on this site, her uncertainty would be a compound of the aleatory and epistemic uncertainties. It would be purely aleatory *if* she knew the distribution of yields in the population of trees that might be grown on that site, but this distribution is *not* known. In particular, its mean is unknown. Uncertainty about features of the population is epistemic and will also contribute to the uncertainty about an individual tree. In statistics, features of populations, such as means and variances, are generally referred to as *parameters* (like the parameters μ, α, τ and σ in the medical example of Section 1.3.1). The theory of statistics is concerned with ways to make inferences about the unknown parameters, using the available data. The things that we wish to ask experts about are very often what statisticians would call parameters. Uncertainty about them is always epistemic because the population is unique, and elicitation is always concerned with the expert's personal probabilities.

There is controversy in the world of statistics about the use of personal probability. The most widely taught theory of statistical inference is the frequentist theory, in which parameters are regarded as unknown but fixed. In frequentist statistics, it is not legitimate to express probabilities about parameters, because only the frequency interpretation of probability is admitted. The rejection of personal probability as a basis for scientific reasoning is one of the differences that distinguishes most followers of frequentist statistics from most advocates of Bayesian statistics, the latter generally embracing personal probability in their methods. However, in the practical elicitation of expert knowledge, this controversy does not arise. The focus of attention in practice is *always* on variables for which there is at least a component of epistemic uncertainty, and expert judgements are therefore always personal probabilities.

In the light of the timber yield example, we can now refer back to the problem in Section 1.2.1 of assessing the probability that I will be killed in a road accident in the next 12 months. It was noted there that the combination of relevant factors – my age and gender, where I live, the kind of car I drive, and so on – make me unique, so that it is no longer possible to ascribe a probability by referring to road accident

data. This is clearly analogous to the uniqueness of the timber company's site. It may be entirely natural to ask about the probability that I will be killed on the road in the next 12 months, but it is not possible to give a frequency interpretation to such a probability. The only sense in which we can discuss it meaningfully is within the personal probability framework. The fact that people, in general, are willing to talk about unique events as having probabilities emphasises the importance of personal probability.

1.3.4 Implications for elicitation

Most people are familiar with probability only in terms of repeatable, random events, and this has important implications for the process of elicitation. If an expert is asked to express her probability for the proposition that the asthma drug will increase FEV_1 by an average of 100 ml or more, or that the mean yield will exceed 45 m^3 per tree, we are asking for a personal probability. In trying to answer the question, she cannot appeal to any experience of repetitions since the events she is being asked about are unique and repetition is impossible. Nevertheless, the familiar ideas of frequency probability are a valuable guide.

First, when explaining to the expert what is needed, it is usual to draw analogies between personal probabilities and frequencies. The expert will be advised that she should give a probability of one-sixth if she has the same strength of belief in the proposition as in throwing a 6 with a die. Well-known frequency probabilities associated with familiar gambling devices such as dice, coins, roulette wheels and cards help the expert assign personal probabilities to one-off propositions.

Second, experience with frequencies of related things may suggest a probability. For instance, the medical expert may know that six out of seven asthma drugs claim to increase FEV_1 by at least 100 ml. This is not really repetition because the drugs are all unique, but it gives the expert a sense of how realistic it is for a new drug to reach that level of effect. In the same way, the forestry expert will use the yields of the tree species in other places to indicate a probability, but must also account for the unique features of the specific site. The knowledge that an expert draws on is often a kind of quasi-repetition, moderated by a judgement of how much the proposition in question is representative of those quasi-repetitions.

1.4 Elicitation and the psychology of judgement

When we talk of 'elicitation', we imply that our respondents have some kind of knowledge or beliefs 'in their heads' and it is our task to devise the right kind of questions to 'extract' this information from them. But is this picture correct? Do people have ready-formed beliefs waiting to be extracted in this way? And even if they do, if such beliefs concern uncertain events or prospects, are they represented subjectively in terms of numerical probabilities? To start answering such questions, we need to go back a bit into history to remind ourselves of the concerns that have guided the development of theory and method in psychology.

1.4.1 Judgement – absolute or relative?

Psychology is, very largely, the scientific study of how human beings think and feel and act on the basis of their thoughts and feelings. So one of the first questions is how we can tell what someone is thinking or feeling. This, of course, was, and still is, a fundamental question of philosophy, but what distinguished the aspirations of early experimental psychologists was a conviction that thoughts and feelings were, in principle, *measurable*. Those who initiated this programme of work in the mid-nineteenth century referred to their field as *psychophysics*. Partly, this signified an intention to bring the methods of the physical sciences to bear on the subject matter of psychology, but it had a more precise sense too. The search for the 'psychophysical law' involved attempting to specify, in exact mathematical terms, a function that described the relationship between 'psychological magnitudes' on the one hand and 'physical magnitudes' on the other.

The context in which this enterprise was undertaken was that of sensory perception. Hence, the 'psychological magnitudes' studied were sensations – of the loudness of tones, the brightness of lights, the length of lines and (particularly in the early days, since such stimuli could be manufactured easily) the perceived heaviness of different weights. Consider the last of these. Participants are presented with a series of small brass cylinders, made to be identical visually but of different actual weights. The 'physical magnitude' of these stimuli is simply their weight in grams and their 'psychological magnitude', is, how heavy they feel. Obviously, these two sets of magnitudes will be related to each other. A weight of 200 g will feel heavier than one of 100 g, but will it feel exactly twice as heavy? And will the difference between these two weights feel the same as that between two weights of 200 and 300 g? The broad answer is no. For many sensory continua – at least those that involve changes in perceived intensity rather than in quality (colour is an example of the latter) – sensitivity to differences is reduced as the stimulus intensity increases.

The so-called Weber–Fechner law (Fechner, 1860) proposed that the 'difference threshold' or 'just noticeable difference' (JND) for any stimulus – the amount by which the magnitude of the stimulus would need to be increased for the difference to be just detectable – is a logarithmic function of the distance of that stimulus above 'absolute threshold' (i.e., the smallest detectable stimulus intensity). It was also assumed that the perceived difference between any pair of stimuli is a direct function of the number of JNDs by which they were separated. This law remained intact for a century until Stevens (1957) proposed its replacement by a power function. Even then, the differences between the predictions of these two versions of the 'psychophysical law' are quite subtle over many ranges of stimulus intensity.

In the course of their search for this precise mathematical function, however, psychophysicists were quickly confronted by another phenomenon. Judgements are highly dependent on context. Perhaps the most common context phenomenon is termed the *contrast effect*. Here is a typical experiment. In a control condition, participants have to lift a series of weights ranging between, say, 200 and 400 g and rate how 'light' or 'heavy' they feel (e.g., on a scale with, say, 11 response

categories labelled 'extremely light' at one end and 'extremely heavy' at the other). In other conditions, an 'anchor' or standard weight (that need not be judged) is added on alternate trials between each of the original variable weights. If this anchor is much heavier than the original series (say 900 g), the remaining weights will be judged lighter; if the anchor is lighter than the original series (say 50 g), they will be judged heavier.

Interestingly, the original reason for introducing such 'anchors' was, as the term suggests, to stabilise participants' use of the rating scale so that the relationship between the physical magnitudes of the stimuli and how they were rated remained relatively unchanged across the course of the experiment. Sometimes these anchors could fall within the range of the stimuli to be judged and sometimes at or beyond the extremes of the stimulus distribution. Sometimes participants were instructed that the anchor or anchors corresponded to a specific category or score on the response scale; sometimes this was left unstated or implicit. Either way, this procedure did generally achieve the desired effect. In other words, the resulting judgements tended to be more consistent in terms of their rank ordering and defensibly interpretable as constituting an interval scale (as required for any test or application of the 'psychophysical law'). However, participants' judgements, while internally consistent, were entirely relative to the context in which they were presented. Although a 350-g weight would always receive a 'heavier' rating than one of 250 g *in the same experimental condition*, there is no guarantee at all that a 350-g weight in one condition would be rated as 'heavier' than a 250-g weight *in a different condition* (as for instance, if the stimuli in the first condition ranged from 200 to 800 g and those in the second condition, from 50 to 400 g). In short, such judgements are always relative, not absolute.

Contrast effects are easy to demonstrate, but less easy to explain definitively. A major ambiguity is whether such effects reflect changes in sensation (or 'psychological magnitude') resulting from perceptual adaptation (as when an indoor room initially seems dark after coming in from bright sunshine) or merely a 'semantic shift' in terms of participants defining descriptive terms such as 'light' and 'heavy' to match the range of stimuli actually presented. According to the latter interpretation, a weight of 400 g would not necessarily *feel* any lighter in the context of a 900 g anchor. It is just that the term 'extremely heavy' now has to allow for weights of at least 900 g, whereas in the control condition, it could be used for weights of just 400 g. Probably both perceptual adaptation and semantic shifts are present in this example, but separating out their effects is enormously difficult with methods such as those described. Of more general significance than debates over the role of perceptual adaptation, however, is the fact that relativity to context pervades all kinds of judgements, not just those involving sensory perception. Judgements of personality, of the seriousness of offences and of political and social attitudes all show similar effects (see, for example, Eiser, 1990; Eiser and Stroebe, 1972).

What are the implications of this piece of history for the elicitation of experts' probabilities? The main messages are, firstly, that subjective perceptions and

sensations are, in principle, measurable – and with some precision – but such measurements can only be interpreted relatively and not absolutely. Secondly, in recognition of such relativity, the psychology of judgement needs to take on issues that extend beyond the original aim of the psychophysicists in relating psychological and physical continua to each other. Simply stated, there is a logical distinction between people's sensations, or subjective representations, and their *descriptions* of such sensations in terms of any response scale. In other words, whereas classical psychophysics was primarily concerned with the relationship between continua of psychological and physical magnitudes, more modern judgement research addresses the relationship between psychological and response continua.

Translating this into the question of eliciting probabilities, we therefore need to remember two basic distinctions. The first distinction, analogous to that between psychological and physical magnitudes, is that between people's subjective representations of how probable things are and the objective events or evidence that provide grounds for such representations. The second distinction, corresponding to that between psychological and response continua, is that between people's subjective representations of how probable things are and the manner in which they express such representations on any given response scale. For example, just because respondents rate something as '90% certain', we cannot assume that they really *mean* 90% rather than 95%, 85% or even 65% or that what they think of as 90% equates to what anyone else would mean by 90%. However, we *can* assume that they mean it is more probable than something else they have rated as '80% certain'. In other words, just because respondents may seem happy to use numerical probabilities to *express* an estimate, we cannot assume that they represent such estimates numerically to themselves (though they may) and, even less, that they have arrived at any given numerical estimate through a process of normatively correct statistical reasoning (though with training they sometimes may).

The argument, then, is that judgements of probability, however elicited, are just that – *judgements*. Knowing what we do about the sensitivity of all judgements to context effects, we should always be wary of interpreting them in absolute rather than relative terms. Nonetheless, judgements of probability at least *appear* more absolute than, say, judgements of heaviness or loudness. This is largely because the meaning of different probability values is well defined (i.e., absolute), is assumed to be well known and, importantly, the scale on which they fall is bounded at both extremes (0 and 1). In comparison, many scales of psychological measurement are not bounded at all, and some (e.g., perceptual judgements) only have a minimum or lower bound (i.e., absolute threshold) but no upper bound. In this sense, probability judgements on a scale of 0 to 1 are already 'anchored' at the two extremes. So does this remove the problem of relativity? Not quite.

The first reason is conceptual. People's perceptions of how probable things are may not correspond to any formal definition of probability. By analogy, we know that the loudness of sounds can be measured in decibels, but asking people (even expert sound engineers) to *estimate* the decibel value of a given sound in no way guarantees that their estimates will be accurate. Eliciting probabilities from experts

is not a straightforward matter, and context effects may be expected to contribute to errors of judgement. The second reason is more empirical and relates to the actual distribution of probabilities which judges are asked to estimate when required to make a series of judgements, for instance, whether the distributions are positively or negatively skewed and how much of the range from 0 to 1 they cover. A well-established principle, related to the contrast effects previously described, was first identified by Parducci (1963) and is termed the *range-frequency compromise*. This involves two tendencies. The first is for judges to use the different regions (or categories) of a response scale to cover broadly equal intervals or fractions of the total range between the smallest and the largest stimulus presented. The second is for judges to use the different regions or categories with broadly equal frequencies. Of course a scale such as 'extremely light' to 'extremely heavy' can be assumed to be more vulnerable to such effects than a probability scale from 0 to 1. However, the influence of such a principle cannot be ruled out in situations where the actual probabilities being judged are heavily skewed. Even an association that is very strong in epidemiological terms (e.g., the conditional probability of contracting lung cancer if one is a smoker, about 0.1) can fall very close to the bottom end of the 0 to 1 scale. If judges are asked to make comparative estimates, on this scale, of various mathematically small probabilities, it could well be the case that they would achieve 'better' differentiation by overestimating some of the less improbable events. Another way in which the range-frequency principle might have an effect could be when judges are asked to provide estimates of distributions, rather than single probabilities. A speculation implied by research in other areas is that, depending on the elicitation procedure, judges may produce distributions that are less skewed, and possibly less peaked, than they should be.

1.4.2 Beyond perception

A possibly less helpful legacy of classical psychophysics has been the tendency to conceive judgement prototypically as a form of perception. It is as though participants are told, "Here's a stimulus, please look at it (or listen to it or touch it) and tell me what you see (or hear or feel)." Sometimes this prototype is quite appropriate, as when a stimulus is physically present and can be directly perceived. But not all judgements are like that. Frequently, we are asked to make judgements of concepts or hypothetical events and consequences. Examples of the latter include judgements of political preference and trust in politicians, elicitation of probabilistic estimates of the outcomes of medical treatment or of the susceptibility of members of a given population to a particular disease. Here, there is no specific thing that has a determinable 'physical magnitude'. Even where the question is one of aleatory uncertainty and there is a correct answer determinable in advance (e.g., the prevalence of disease D in population P), this fact is not known as such to the respondents (or there would be no point in asking for their judgements). So they are not 'perceiving' the prevalence and then describing their 'perception'. They are thinking about the problem, forming some notion of what the prevalence

might be and then expressing this notion in terms of the particular response format presented to them by the researcher. Such 'thinking about' involves a much more deliberative and effortful process than that typically implied by more automatic perceptual responses. The challenge, then, is to determine what this 'thinking about' comprises.

The starting point for an analysis of this problem is to realise that this 'thinking about' is a process that is set in motion by the researcher's interrogation. This process leads to the respondent arriving at a notion of a possible answer and then translating this notion into the response language provided. Within this process, there is lots of room for conceptual slippage. Does the respondent understand the researcher's question in exactly the way the researcher intended? Does the respondent then continue to think only (and wholly) about the question asked or does he or she start thinking about other associations that are not strictly part of the problem itself, while selectively failing to properly attend to all the features of the problem that are relevant? (Note the discussion of cognitive heuristics in Chapter 3.) And once the respondent has arrived at a notion of how to answer the question, will the answer then offered be interpreted by the researcher exactly as intended?

It is now acknowledged that an extremely important influence on this process of 'thinking about' is memory. There are a number of reasons why memory is so important for judgement. One reason is that memory is selective. Suppose that our expert is asked to estimate the prevalence of condition or disease D in population P (e.g., the proportion of children under five in an East African state who are likely to die from malnutrition). Our respondent (who, let us assume, has never been to Africa) will try to remember relevant sources of evidence, from news reports, from articles in professional journals, from conversations with colleagues, and so on. Some of this information may be easier to recall because of the vividness of news reports. In other words, the information on which the respondent's judgement is based may be only a sub-sample, or selection, of what is potentially relevant and available. Judgements can also be influenced by memory associations, that is, by thoughts triggered by irrelevant (or non-diagnostic) features of the problem, or by over-generalisation from inferences based on the membership of a broader category.

To continue our example of malnutrition in Africa, let us suppose that, whereas many states in that part of the world have had years of poor harvests, the state in question has remained relatively prosperous. If so (and particularly if this has not been considered newsworthy by western media), our respondent may overestimate the mortality rate of the specific country as a consequence of a broadly accurate but undifferentiated association between Africa, poverty and famine. Thus, through the combined effect of selectivity and associative memory, our respondent may fail to recall some useful information, and let other, misleading or less relevant, information intrude. The point is that our respondent's prevalence estimate is not something ready formed and 'sitting there' in some memory store just waiting to

be retrieved. It is something constructed from the ideas and associations that come to mind while the respondent thinks about how to answer the question.

In short, memory can involve the reconstruction of meaning and not simply the recall of facts or events. Related to this is the fact that we frequently have little control over what memory associations do come to mind and little insight into why we remember what we do or what has triggered particular thoughts. There is a considerable body of recent literature on how 'automatic' memory associations can influence attitudes, judgements and decisions, with many of these processes occurring below the level of conscious awareness. See, for example, Bargh et al. (1996), Bargh and Ferguson (2000) and Fazio (2001).

1.4.3 Implications for elicitation

We can draw from this generic psychological research some important conclusions about the process of eliciting experts' probabilities. First, an elicited probability is a judgement, and we should expect in principle that generic findings about judgements will apply. In particular, research has shown that judgements of stimuli such as weights are intrinsically relative. Even when anchors are provided, changing the anchors changes the context and is likely to influence the respondents' judgements. However, it has already been remarked that probability judgements are different with respect to having natural limits of 0 and 1. These act as absolute and unvarying anchors. Furthermore, it is usual in eliciting probabilities to give the expert additional absolute anchors in the form of reference events having specific probabilities. For instance, the probability of 0.5 is explained as corresponding to a proposition that is equally likely to be true or false, equivalent to the event of drawing a red ball from a bag containing one red and one white ball. These practical measures should mean that probability judgements suffer much less from relativity effects than those found in experiments using judgements of other stimuli.

We should, nevertheless, expect to find that the ways that people make such judgements lead to biases, such as through the range-frequency compromise. A number of such bias-inducing heuristics are discussed in Chapter 3.

It is also important to recognise that experts construct probability judgements in response to the stimulus of questioning; their probabilities are not pre-formed values simply waiting to be expressed. The role of memory in this process, and the effect that different forms of questions can have on which memories are accessed, also underlies some of the sources of bias in probability judgements that are discussed in Chapter 3.

1.5 Of what use are such judgements?

The purpose of elicitation is to obtain a formal expression of the expert's knowledge regarding an uncertain quantity (or quantities). Usually, the resulting probability distribution will be used as part of some analysis (for instance a risk analysis) or to

aid in making a decision. It is therefore important to consider the extent to which probability judgements elicited from an expert have the status and interpretation that is expected of them in these applications, particularly bearing in mind the preceding discussion about how those judgements are constructed in practice.

1.5.1 Normative theories of probability

The probabilities that we wish to elicit (and which statisticians have often implicitly assumed are indeed elicited) are those that are implied by normative theories of decision-making under uncertainty. The relevant theory was first fully developed by Savage (1954) and further expounded in the classic textbook of DeGroot (1970). According to this theory, in order to make decisions which satisfy some natural axioms, persons must behave as if they (a) have a probability distribution for all relevant uncertain quantities, (b) have a utility function expressing the value of making any given decision conditional on the true values of those quantities, and (c) choose an optimal decision as the one that maximises expected utility. In principle, the expert's probability distribution can be revealed by offering enough different decision options and rewards and then observing what decisions she makes.

In another seminal work, de Finetti (1974) formally defined probability to be the decision made in response to a quadratic scoring rule (see Section 8.2). Specifically, suppose that the expert states her probability for an event E to be q. Then when it is determined whether E does occur, she receives a reward $1-(1-q)^2$ if E occurs and $1-q^2$ if it does not (in some appropriate monetary units). If the expert actually judges the probability for E to be p, then her expected reward in stating it to be q is

$$\begin{aligned} p\{1-(1-q)^2\}+(1-p)\{1-q^2\} &= 1-p(1-q)^2-(1-p)q^2 \\ &= 1-p(1-p)-(p-q)^2. \end{aligned} \tag{1.1}$$

This expected reward is maximised by the expert stating a probability, q, that is equal to her actual assessment p. So de Finetti's scheme encourages the expert to assess her probability accurately. However, it assumes that she is able to balance probabilities and rewards appropriately. Under this assumption, it is proved that the expert's probabilities will behave according to the laws of probability theory and will be appropriate for use in subsequent analyses or decision-making.

These normative theories state that probabilities are the uniquely scientific way to represent uncertainty. Furthermore, probabilities defined according to such theories are what is needed for use in applications such as risk analysis or decision-making.

1.5.2 Coherence

Whatever interpretation we place on probability, frequentist or personal, it is agreed that probabilities should obey the laws and theorems of probability theory (such as

the Addition Law of Section 1.2.1 and the Multiplication Law of Section 1.2.4). A set of probability judgements that follow all these laws and theorems are called *coherent* (see Section 8.3 for more discussion of coherence). In developing their theories of personal probability, Savage, DeGroot and de Finetti took care to show that the probabilities that would be obtained must be coherent. For example, de Finetti (1974) shows that a person who assigns a series of probabilities according to the reward scheme illustrated in (1.1) will necessarily expect to obtain a lower reward if her probabilities are not coherent. The assumption that she is able to combine accurately the probabilities and rewards (and that she wishes to maximise her expected reward) would imply that she would not make this mistake, and hence her probabilities must be coherent.

In practice, however, we know that people do assign probabilities non-coherently. They make errors of judgement that are assumed not to happen in the normative theories. This raises the fundamental question of the nature of elicited probability judgements and the extent to which they can be treated as having the interpretation that is required for practical risk assessment and decision-making.

1.5.3 Do elicited probabilities have the desired interpretation?

Both de Finetti and Savage considered the process of obtaining expert responses to choices with rewards such as (1.1) to be elicitation; see, for instance, Savage (1971). In general, though, psychologists would regard such choices as cognitively more complex than asking directly for an assessment of probability. It seems unlikely that experts would perform better in such tasks. If, in order to determine q, the expert follows the above reasoning and decides that her answer ought to be her probability p, then she still has to determine p with all the difficulties outlined above and in Chapters 3 and 4. If a less analytical approach is used, without explicit assessment of probability but with the expert making a more intuitive choice in the face of the reward scheme, then this seems likely to lead to a less accurate judgement through imperfect appreciation of the implications of the reward scheme.

Research and the practice of elicitation have since concentrated on the direct elicitation of probabilities. However, this compounds the question of what interpretation we can place on the elicited probabilities. If they have not been obtained according to their formal constructions in the theories of Savage, DeGroot or de Finetti, and if they may, in practice, be non-coherent, what status do they have?

Winkler (1967, p. 778) writes

> "The assessor has no built-in prior distribution that is there for the taking. That is, there is no 'true' prior distribution. Rather, the assessor has certain prior knowledge which is not easy to express quantitatively without careful thought. An elicitation technique used by the statistician does not elicit a 'true' prior distribution, but in a sense helps to draw

> out an assessment of a prior distribution from the prior knowledge. Different techniques may produce different distributions because the method of questioning may have some effect on the way the problem is viewed."

Winkler seems to regard the different distributions that result from different elicitation techniques as equally valid, but this would deny the considerable research in the psychology literature (see Chapter 4) that demonstrates how some forms of questioning lead the expert to view the problem inappropriately, in the sense that they do not utilise the available information fully and accurately.

On the other hand, O'Hagan (1988) explicitly defines 'true' probabilities as those that would result if the expert were capable of perfectly accurate assessments of her own beliefs. He shows that such 'true' probabilities will satisfy the coherence requirements of the normative theories. O'Hagan regards different 'stated' probabilities, that might result from different elicitation methods, as more or less inaccurate attempts to specify the expert's underlying 'true' probabilities.

It is clear that whether or not we believe that experts' knowledge is representable by unique, *true* probability distributions there are ways in which the expert might give poorly judged assessments. So not all elicitation techniques will lead to equally valid results. It is important that the expert should view the problem from as complete a perspective as possible, utilising all the relevant information in an unbiased way. If this were achievable, taking care, in particular, to avoid the sources of poor judgement and bias that have been identified by psychological research, then the elicited probabilities and distributions would be coherent. We could call such a set of probabilities or probability distributions *good*.

In practice, an elicited probability distribution can be seen as an approximation to such a 'good' distribution. O'Hagan (1988) and the earlier developers of personal probability implicitly assumed that there would be a unique 'good' distribution, which can then be called 'true', but this is an open question. Are people's probabilities different only because no two people have identical knowledge and experiences, or if we could ever find two people who are identical in this respect might they legitimately have different probabilities? In a similar vein, if two elicitation techniques were both so perfect that they yielded 'good' distributions, could they, nevertheless, produce different distributions? There are theories of probability that take an 'objective' view that there is a unique probability distribution associated with any state of knowledge. They are associated, in particular, with the debate about probabilistic representation of ignorance (Jeffreys, 1967; Kass and Wasserman, 1996).

We take the view that the purpose of elicitation is to represent an expert's knowledge and beliefs accurately in the form of a 'good' probability distribution. In later chapters, we may refer to this as a true distribution, but the reader should be aware that this does not necessarily imply that the true distribution is unique.

1.6 Conclusions

Each chapter of this book will conclude with a summary of its findings relating to good elicitation practice and areas of research need. These are collected in Chapters 11 and 12.

1.6.1 Elicitation practice

- The distinction between aleatory and epistemic uncertainty is important for elicitation practice. Elicitation usually focuses on uncertainties that are either purely epistemic or have an epistemic component. However, people are most familiar with the concepts of probability in the context of aleatory uncertainties.
- It is important to remember that elicited statements of probability are judgements made in response to the facilitator's questions, not pre-formed quantifications of pre-analysed beliefs. The psychophysics literature suggests that all such judgements are intrinsically relative.
- The range-frequency compromise suggests that in some situations experts will tend to distribute their elicited probabilities evenly over (the whole or part of) the probability scale.
- Elicited probabilities may suffer from biases and non-coherence in practice, but the goal of elicitation is to represent the expert's knowledge and beliefs as accurately as possible.

1.6.2 Research questions

- To what extent does the existence of an absolute scale (0 to 1) for probabilities, and the way that training usually gives the expert other anchors or landmarks on the scale, allow absolute (rather than relative) judgements?
- What are the implications of the range-frequency compromise in the context of probability elicitation?
- Does elicitation using proper scoring schemes (as propounded by Savage and others) lead to less accurate assessments than the direct elicitation of probabilities?