Uncertain Judgements: Eliciting Experts' Probabilities

Anthony O'Hagan[*], Caitlin E. Buck[*], Alireza Daneshkhah[*], J. Richard Eiser[*], Paul H. Garthwaite[†], David J. Jenkinson[†], Jeremy E. Oakley[*] and Tim Rakow[‡]

[*] *University of Sheffield, UK* [†] *The Open University, UK*

[‡] *University of Essex, UK*

John Wiley & Sons, Ltd

Uncertain Judgements

Statistics in Practice

Statistics in Practice is an important international series of texts, which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. The feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Uncertain Judgements: Eliciting Experts' Probabilities

Anthony O'Hagan[*], Caitlin E. Buck[*], Alireza Daneshkhah[*], J. Richard Eiser[*], Paul H. Garthwaite[†], David J. Jenkinson[†], Jeremy E. Oakley[*] and Tim Rakow[‡]
[*]*University of Sheffield, UK* [†]*The Open University, UK*
[‡]*University of Essex, UK*

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-02999-2 (HB)
ISBN-10: 0-470-02999-4 (HB)

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface

This book arises from a multi-disciplinary research project commissioned by the R&D Research Methodology Programme of the National Health Service in the United Kingdom. The BEEP (Bayesian Elicitation of Experts' Probabilities) project brings together expertise in statistics, psychology and health. Although based in the Department of Probability and Statistics at the University of Sheffield, BEEP includes researchers from two other departments at Sheffield and from the Open University and the universities of Essex and Leeds. The first task within the BEEP research programme was to produce an authoritative review of the diverse literature on elicitation. As our perspective over this broad field grew, it became clear that what we were compiling was more than a review. The reader will find herein an extensive coverage of the field of elicitation, with very many references to the literature, but will also find commentary and synthesis. The result is a book that sets out our view of what elicitation is, what constitutes the best current practice in elicitation and the outstanding areas where research is needed, and it also provides detailed citations of the existing literature.

It has been written by the following BEEP team members:

- University of Sheffield, Department of Probability and Statistics: Professor Tony O'Hagan, Drs Caitlin Buck, Alireza Daneshkhah and Jeremy Oakley.
- University of Sheffield, Department of Psychology: Professor Dick Eiser.
- Open University: Professor Paul Garthwaite and Mr David Jenkinson.
- University of Essex: Dr Tim Rakow.

The authors have also benefited from valuable comments and contributions from other team members and the BEEP international advisory group. We are particularly grateful for contributions from Dr Gaëlle Villejoubert (University of Leeds) and Professors Roger Cooke (Delft University of Technology), Baruch Fischhoff (Carnegie Mellon University), Nigel Harvey (University College London), Colin Howson (London School of Economics), Jay Kadane (Carnegie Mellon University), Larry Phillips (London School of Economics) and Bob Winkler (Duke University).

Further information about BEEP can be obtained from the project website <www.shef.ac.uk/beep/>

Scope of the book

A number of reviews of elicitation have been published, including Chaloner (1996), Cooke (1991), Garthwaite et al. (2005), Hogarth (1987), Kadane and Wolfson (1998) and Wallsten and Budescu (1983). With the exception of Garthwaite et al. (2005), where the emphasis is on the statistics literature, these are all now out of date. We have built on these works and made use of our own knowledge of the literature, but have also carried out extensive new searches.

- We searched the ISI Science, Social Science and Arts and Humanities Citation Indices under the terms 'probability assessment', 'subjective probability', 'elicitation' and 'expert opinion'. More than 5000 references were found and abstracts were checked where available. More than 1000 articles were identified as relevant and investigated further.

- In keeping with the BEEP project's funding source, we made particular efforts to identify articles in the medical literature, searching the MEDLINE database under the same four terms. More than 8000 references were found and some 900 were identified for further investigation.

- A hand search of *Organizational Behavior and Human Decision Processes* between January 1990 and January 2004 was also conducted. There is a considerable amount of material in this journal that relates to the psychological issues around human judgement. Approximately 100 further papers were identified in this way as being relevant to the book.

- Of the 2000 or so references identified in these ways for further investigation, careful reading of abstracts led to the selection of more than 400, whose full text was retrieved and read.

- We asked each of our international advisory panel members to give us a short list of seminal papers in their own area of expertise; they also checked the final content of the book.

In this way, we hope that this book will be comprehensive in its scope, authoritative and up to date. Inevitably, there will be omissions and the discussion and emphasis will reflect the judgements of the authors about which ideas are the most useful and important.

Target audience

We hope that the book will be found useful by a wide range of people. Those who need to take decisions in the context of substantial uncertainty and risk know that they must often rely on expert judgement. Decision makers should find in this book a valuable introduction to how the elicitation of those judgements in probabilistic form can be achieved, and an overview of the current state of the art. Researchers

in psychology (particularly the psychology of judgement and decision-making), decision theory, risk assessment and statistics (particularly Bayesian statistics) should find valuable synthesis of parts of their own research areas, with indications of the gaps that we feel need to be filled by new research. More importantly, they will see how the work in their own area relates to research in other disciplines and will gain a more complete appreciation of the field of elicitation.

With such a diverse readership in mind, there is a risk that the technical jargon of one discipline will be impenetrable to those whose background is in another. We have tried to introduce ideas in a simple manner wherever possible. We have also provided an extensive Glossary, covering the major concepts in both statistics and behavioural psychology. Even the reader who skips over the most technical parts of unfamiliar subjects should find enough less technical discussion to obtain a general understanding.

Outline of the book

The book is structured as follows:

- Chapter 1 sets out some fundamental concepts regarding elicitation, probability and expert judgement.
- Chapter 2 outlines the elicitation process in general terms, highlighting the differing perspectives of statistical and psychological research in elicitation.
- Chapters 3 and 4 are concerned with psychological theories and experimental evidence about expert judgement of uncertainty and with how these findings relate to the practical elicitation of probabilities.
- Chapters 5 to 7 deal with eliciting probability distributions, primarily from a statistical perspective.
- Chapter 8 discusses ways to evaluate the accuracy of elicited probabilities and distributions.
- Chapter 9 deals with issues when eliciting from multiple experts.
- Chapter 10 reviews a selection of applications, to give a flavour of the variety of methods that are actually in use.
- Chapters 11 and 12 collect together the main findings in Chapters 2 to 9, with regard to best elicitation practice and areas where further research is needed.
- The book ends with an extensive Bibliography of cited references, and a Glossary that explains some common terms in psychology and statistics.

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

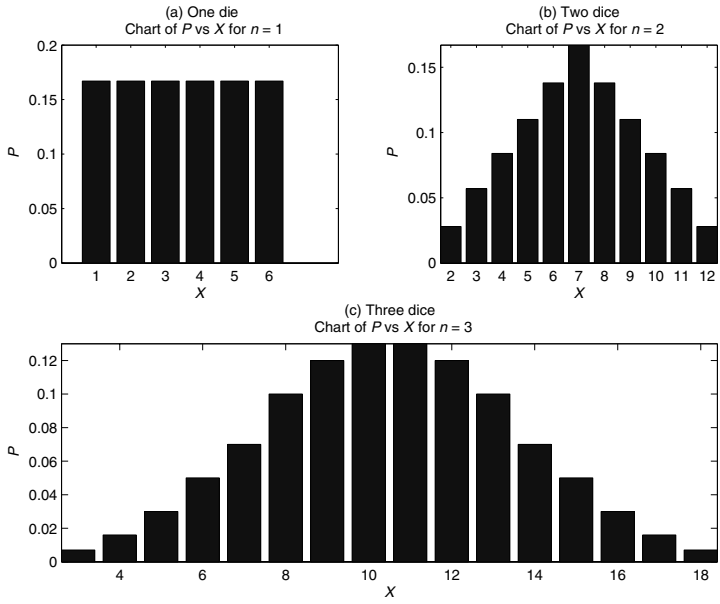

Figure 1.1: Probability mass functions for total score on n dice.

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

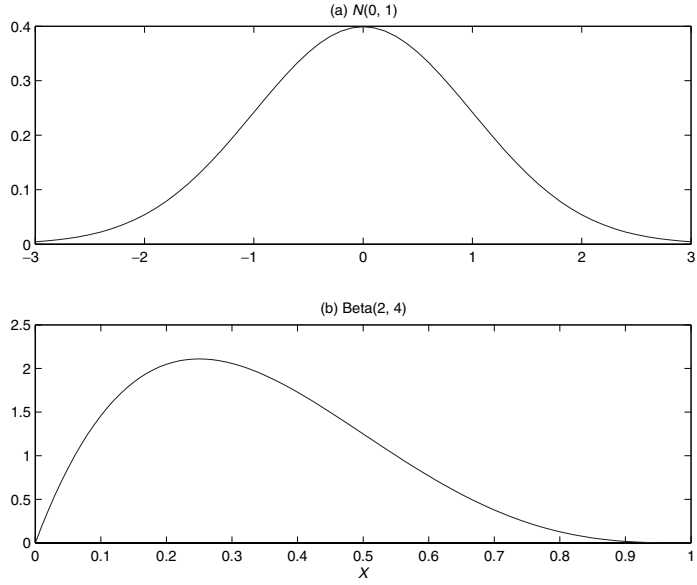

Figure 1.2: Two probability density functions.

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.