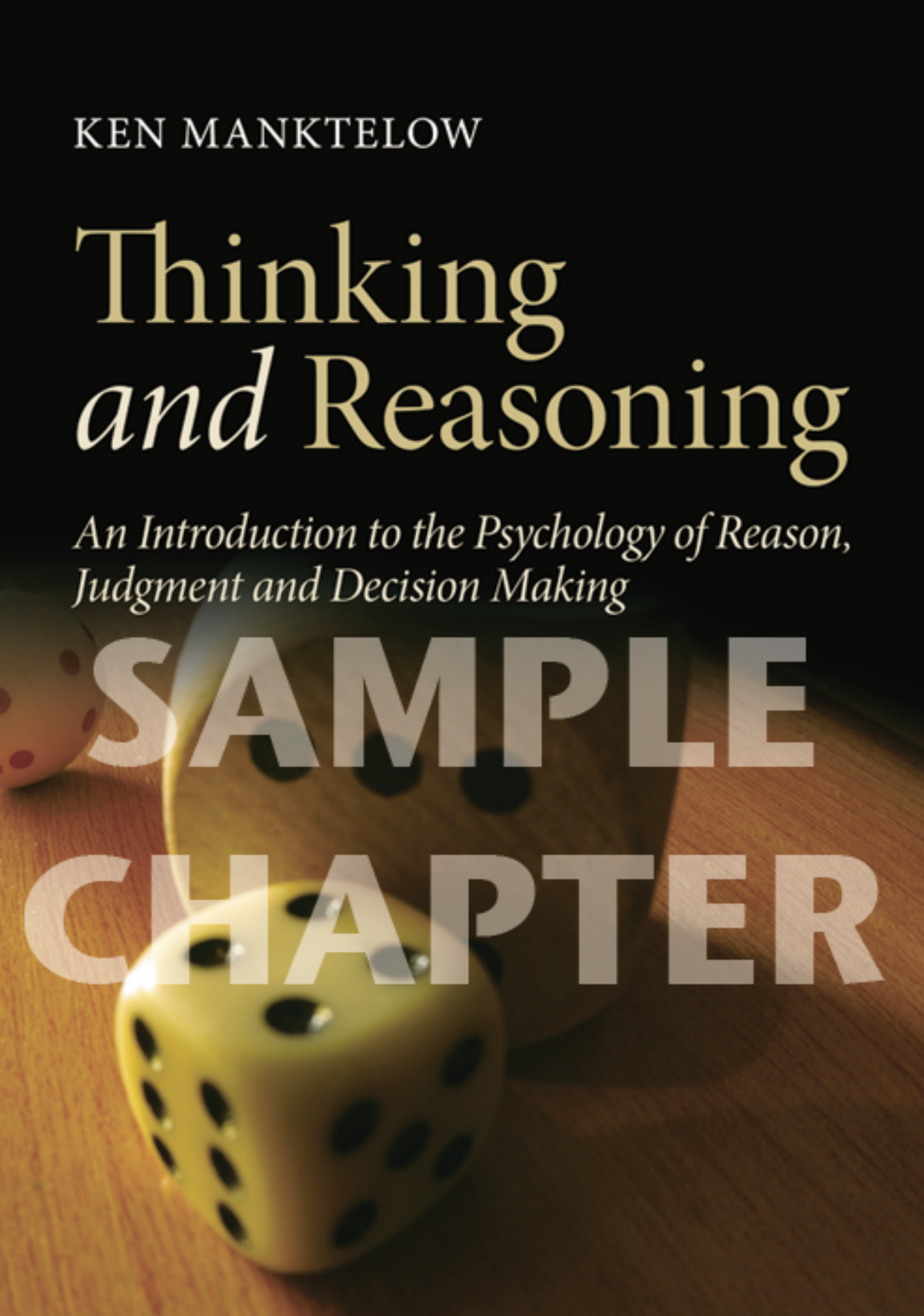
KEN MANKTELOW

Thinking *and* Reasoning

An Introduction to the Psychology of Reason, Judgment and Decision Making



First published 2012
by Psychology Press
27 Church Road, Hove, East Sussex BN3 2FA

Simultaneously published in the USA and Canada
by Psychology Press
711 Third Avenue, New York, NY 10017

[www.psypress.com]

Psychology Press is an imprint of the Taylor & Francis Group, an informa business

Typeset in Century Old Style and Futura by RefineCatch Ltd, Bungay, Suffolk

Cover design by Andrew Ward

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Manktelow, K. I., 1952–
Thinking and reasoning : an introduction to the psychology of reason, judgment and decision making / Ken Manktelow.
p. cm.
Includes bibliographical references and index.

1. Reasoning (Psychology)
2. Thought and thinking.
3. Cognition. 4. Decision making.
I. Title.
BF442.M354 2012
153.4'2–dc23

2011031284

ISBN: 978-1-84169-740-6 (hbk)
ISBN: 978-1-84169-741-3 (pbk)
ISBN: 978-0-203-11546-6 (ebk)

Contents

Chapter 1

Judging and thinking about probability

http://www.psypress.com/thinking-and-reasoning-9781841697413

If you have read the Preface, you will have been cued into the importance of probability in how we now understand and explain thinking and reasoning. We therefore open with this central topic: it forms the basis of explanation across the spectrum of thought. In this chapter, we shall look at research that assesses how people judge probability directly. You may have come across the word in statistics classes, and already be stifling a yawn. If so, wake up: probability is far from being just a dry technical matter best left in the classroom. Everyone judges probability, and does so countless times every day. You might have wondered today whether it is going to rain, how likely you are to fall victim to the latest flu outbreak, whether your friend will be in her usual place at lunchtime, and so on.

How do we make these judgments? We can make a comparison between *normative* systems, which tell us what we ought to think, and *descriptive* data on how people actually do think. To get you going in doing this, here are some questions devised by my statistical alter ego, Dr Horatio Scale. Answering them will raise the probability that you will appreciate the material in the rest of this chapter. We shall look at the answers in the next section.

1 a What is the probability of drawing the ace of spades from a fair deck of cards?
  b What is the probability of drawing an ace of any suit?
  c What is the probability of drawing an ace or a king?
  d What is the probability of drawing an ace and then a king?

2 You are about to roll two dice. What is the chance that you will get 'snake eyes' (double 1)?

3 a What is the chance that you will win the jackpot in the National Lottery this week?
  b What is the chance that you will win any prize at all this week?

  (The British lottery lets you choose six numbers from 1 to 49. You win the jackpot if all your six numbers are drawn; you win lesser prizes if three, four or five of your numbers are drawn.)

4 Yesterday, the weather forecaster said that there was a 30% chance of rain today, and today it rained. Was she right or wrong?

5 What is the chance that a live specimen of the Loch Ness Monster will be found?

6 Who is more likely to be the victim of a street robbery, a young man or an old lady?

7 Think about the area where you live. Are there more dogs or cats in the neighbourhood?

Defining probability

The phrase 'normative systems', plural, was used above because even at the formal level, probability means different things to different people. It is one of the puzzles of history that formal theories of probability were only developed comparatively

recently, since the mid-17th century. Their original motivations were quite practical, due to the need to have accurate ways of calculating the odds in gambling, investment and insurance. This early history is recounted by Gigerenzer, Swijtink, Porter, Daston, Beatty and Krüger (1989) and by Gillies (2000). Gillies solves the historical puzzle by pointing to the use of primitive technology by the ancient Greeks when gambling, such as bones instead of accurately machined dice, and to their lack of efficient mathematical symbol systems for making the necessary calculations – think of trying to work out odds using Roman numerals. All four of the formal definitions of probability that are still referred to have been developed since the early 20th century. Here they are.

Logical possibility

Probability as logical possibility really only applies to objectively unbiased situations such as true games of chance, where there is a set of equally probable alternatives. We have to assume that this is the case when working out the odds, but it is hard to maintain this stipulation in real life. This is behind Gillies' explanation of why the ancient Greeks and Romans could not develop a theory of probability from their own games of 'dice' made from animal bones: these have uneven sides that are consequently not equally likely to turn up.

To see how we can work out odds using logical possibility, let us take Dr Scale's first question, and assume that we are dealing with a properly shuffled, fair deck of cards, so that when we draw one from it, its chances are the same as any other's. There are 52 cards in the deck, only one of which is the ace of spades, so its odds are 1:52. Odds are often expressed in percentages: in this case, it is about 1.92%. Probability in mathematics and statistics is usually given as a decimal number in the range between 0 and 1: in this case, it is .0192. An ace of any suit? There are four of them, so we can work out 4:52 (i.e. 1:13) in the same way, or multiply .0192 by four to obtain the answer: .077.

The odds of an ace *or* a king are the odds of each added together: there is one of each in each suit of 13 cards, so the joint odds are 2:13, or .154. Question 1d is a bit more complicated. It introduces us to the idea of *conditional probability*, because we need to calculate the probability of a king *given* that you have drawn an ace. Each has the probability .077, and you obtain the conditional probability by multiplying them together, which gives the result of just under .006, i.e. 6:1000 – a very slim chance.

Now you can address Dr Scale's third question, the lottery odds. With 49 numbers to choose from, the odds of the first are clearly 1:49. This number is no longer available when you come to choose the second number, so its odds are 1:48, and so on. Sticking with just these two for a moment, the odds of your first one and then your second one coming up can be worked out just as with the ace and king: 1:49 × 1:48, or .0204 × .0208 = just over .0004. But with the lottery, the order in which the numbers are drawn does not matter, so we have to multiply that number by the number of orders in which these two numbers could be drawn. That is given by the factorial of the number of cases, in this case 2 × 1 = 2. This number replaces the 1 in the odds ratio, which now becomes 2:49 × 48. You simply expand this

procedure to work out the chances of 6 numbers out of 49 being drawn in any order. If you want to do this yourself, it is probably easiest to use the numbers in the following form and then use cancelling, otherwise the long strings of zeroes that will result if you use the decimal notation will boggle your calculator:

$$\frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{49 \times 48 \times 47 \times 46 \times 45 \times 44}$$

The top line is factorial 6, for the number of orders in which six numbers could appear. The resulting odds are approximately 1:13.98 million. So if you play once a week, you can expect to win the jackpot about once every quarter of a million years.

Question 3b asked about the odds of winning any prize at all. You can work out the odds of three, four or five of your numbers coming up in exactly the same way as just described. You don't just add all these results together, however, because there are numerous ways in which, say, three numbers from the six you have chosen can come up. So you have to multiply each odds result by this number, and then add them all up. In case you don't feel up to it, I can tell you that the odds of winning any prize in any one week are around 1:57. So a regular player should get a prize about once a year, although this will almost certainly be the lowest prize (the odds of three of your numbers coming up are about 1:54). I shall deal with whether, in the face of these odds, playing the lottery is a good decision in Chapter 8.

Frequency

Frequency theory has roots that go back to the 19th century, but it was developed in most detail in the mid-20th, and has been highly influential ever since. In psychology, this influence has been even more recent, as we shall see later in this chapter. People who adopt frequency theory are called *frequentists*, and they regard probability as the proportion of times an event occurs out of all the occasions it could have occurred, known as the *collective*. There are two kinds of collectives, referred to by an early frequentist, von Mises (1950), as *mass phenomena* or *repetitive events*.

Dr Scale's sixth question can be taken as a frequency question about a mass phenomenon: you could count up the number of street robberies and look at the proportions where the victims were old ladies and young men, and see which was the greater. Games of chance, such as coin flips, dice and lotteries, can also be analysed in terms of frequencies: these are repetitive events. Instead of working out the odds mathematically, you could count up the number of jackpot winners as the proportion of players, for instance. Over time, the results of the two probabilities, frequency and odds, should converge; if they don't, you have detected a bias. Gamblers can be exquisitely sensitive to these biases: Gillies (2000) tells us about a 17th century nobleman whose was able, through his extensive experience with dice games, to detect the difference between odds of .5000 and .4914, a difference of 1.7%. Question 7 (cats and dogs) is also a frequency question, again about a mass phenomenon. Beware: this question has a catch, which we shall return to later in the chapter.

Propensity

For a true frequentist, it makes no sense to ask for the probability of a single event such as how likely you are to win a game of chance, because the only objective probabilities are frequencies – things that have actually happened – and frequencies cannot be derived from one-off observations or events that have not yet occurred. Nor can frequencies have any 'power' to influence a single observation: suppose there are twice as many dogs as cats in your area, does that fact determine in any way the species of the next pet you see? How could it? But everyone has a strong feeling that we can give such odds: we do feel that the next animal is more likely to be a dog than a cat. This clash of intuitions was addressed by the celebrated philosopher of science, Karl Popper (1959a), who introduced the *propensity theory*. He needed to do so because of single events in physics, such as those predicted by quantum mechanics. The probabilities of such events must, he thought, be objective, but they could not be based on frequencies.

Dr Scale's second question is about propensity: note that it asks you specifically about a single event, the next throw, whereas Question 1 was more general. Popper used a dice game example when he tried to solve the intuitive riddle just described (i.e. an example involving logical possibility). Consider two dice, one fair and one weighted so that it is biased in favour of showing a 6 when tossed: it tends to show a 6 on one-third of the throws (i.e. twice as often as it would if unbiased). Suppose we have a long sequence of dice throws, most involving the biased die with occasional throws of the fair one. Take one of these fair tosses. What is the probability that it will show a 6? Popper argued that a frequentist would have to say 1:3, because the 'collective' set of throws has produced this frequency. However, you and I know that it must be 1:6, because we are talking about the fair die, and fair dice will tend to show each of their six sides with the same frequency.

Popper's solution was to appeal to the difference in causal mechanisms embodied in the two dice to resolve this paradox: the collective really consists of two sub-collectives that have been produced by two different generating conditions. The biased die has the *propensity* to show a 6 more often than a fair die does because of its different causal properties. (We shall look in detail at causal thinking in Chapter 4.)

There are problems with propensity theories (others have come after Popper), one of which is that invoking causal conditions just replaces one set of problems with another. In the real world, it can be very difficult to produce the same generating conditions on different occasions, as all psychology students know from the discussion of confounding variables in their methodology courses. If it is not realistically possible to repeat these conditions, then is an objective single-event probability therefore also not possible? And if so, what is the alternative?

Degree of belief

We have strong intuitions about the probability of single events, so the alternative to objective probabilities must be subjective probabilities, or *degrees of belief*.

Look at Dr Scale's fifth question. If you have heard of the Loch Ness Monster, you will have some view as to how likely it is that it really exists. However, this cannot be based on a logical possibility, nor can it be based on a frequency: there is not a 'mass' of equivalent Scottish lochs, some with monsters and some without. Perhaps you can even assign a number to your belief, and perhaps you use objective facts in doing so, to do with the known biology and ecology of the species that you presume the beast to be. But your degree of belief can only be subjective.

Now look at Dr Scale's fourth question, about the weather forecast. Weather presenters often use numbers in this way, but what do they mean? Once again, it is hard to see how this can be a logical possibility: weather is not a series of random, equivalent events, even in Britain. Could it then be a frequency? If so, what of? This particular date in history, or days when there has been a weather pattern like this? Baron (2008) uses weather forecasts as an example by which we can assess how well someone's probabilistic beliefs are *calibrated*. That is, if someone says that there is a 30% chance of rain, and it rains on 30% of days when she says this, then her judgment is well calibrated (this is another kind of frequency). This may be useful information about the climate, but it is not a useful attitude to weather forecasting: we want to know whether to cancel *today's* picnic, a single event. And if the forecaster said 30%, and it rained today, isn't she more wrong than right? She implied that there was a 70% chance that it would not rain. Thus you can be well calibrated in frequency terms but hopeless at predicting single events.

Confusion over the use of percentages like this was addressed by Gigerenzer, Hertwig, van den Broek, Fasolo and Katsikopoulos (2005), following up an earlier study by Murphy, Lichtenstein, Fischhoff and Winkler (1980). First of all, they set the normative meaning of this figure: that there will be rain on 30% of days where this forecast figure is given. So they are adopting the frequentist approach. They tested people's understanding of the figure in five different countries, varying according to how long percentage forecasts had been broadcast. The range was from almost 40 years (New York, USA: they were introduced there in 1965) to never (Athens, Greece). The prediction was that the degree of 'normative' understanding would be correlated with length of usage of percentage forecasts. It was. Alternative interpretations produced by participants were that the 30% figure meant that it would rain for 30% of the time, or across 30% of the region. Note, by the way, that even among the New Yorkers about one-third did not give the 'days' interpretation. And keep in mind that people listen to weather forecasts to find out about particular days, not climatic patterns. Gigerenzer et al. urge that forecasters be clear about the *reference class* when giving numerical probabilities. This is an important issue that we shall return to in a short while.

By the way, my degree of belief in the Loch Ness Monster is close to zero. If, as most of its publicity says, it is a plesiosaur (a large aquatic reptile of a kind that existed at the end of the Cretaceous period, 65 million years ago), we would have to have a breeding population. They were air breathers – heads would be bobbing up all over the place. They would not be hard to spot.

Of course, our beliefs in rain or monsters can be changed if we encounter some new evidence. You go to bed with a subjective degree of belief in rain tomorrow at .3, and wake up to black skies and the rumble of thunder: that will cause you to revise it.

Belief revision: Bayes' rule

The subjective view of probability opens up a range of theoretical possibilities, which all come under the heading of the *Bayesian* approach to cognition. It is hard to overestimate the influence of this perspective at the present time (see Chater & Oaksford, 2008, for a recent survey), and we shall see it applied to theories of reasoning in later chapters. Most of the rest of this one will be concerned with Bayesian matters in some form or other. The word 'Bayesian' comes from Thomas Bayes, an 18th century English clergyman who laid the foundations for this area in a paper published after his death (Bayes & Price, 1763/1970). It is one of the most influential papers in all of science.

To illustrate Bayesian belief revision, Dr Scale has another question for you:

8 Inspector Diesel and his sidekick, Sergeant Roscoe, are on the trail of notorious Dick Nastardley, who is on the run. They know that he is often to be found in his local pub, the Ham & Pickle, on Friday nights – about 80% of the time, they understand.

It is Friday, and Diesel and Roscoe are outside the Ham trying to see whether Dick is inside. They can only see half the bar, and he is not in that half. So, what is Diesel's estimate of the probability that if he and Roscoe raid the pub, they will find their man and nab Dick?

Bayes' rule produces a precise numerical estimate of this kind of probability. It enables us to compute the probability of a hypothesis when given some evidence: a conditional probability. In doing so, we start off with some *prior knowledge* that the hypothesis is true, before the evidence comes in. This can be combined with the *likelihood* of the evidence, given this hypothesis and any alternative hypotheses. These numbers can be combined to derive an estimate of the *posterior probability* that the hypothesis is true, given the evidence.

Question 8 provides you with all the information you need to compute Diesel's degree of belief that Dick is in the bar. The prior probability is his existing belief (based on a frequency) that he is in the bar – 80%, or .8. The alternative hypothesis – there is only one in this case – is that he is not there: Diesel holds this at .2. Now comes the evidence, or data: Dick is not in the visible half of the bar. If he were in the Ham & Pickle, the probability that he would not be visible is obviously .5, because he could be in the half that is visible or the half that is out of sight; if he were not there, this probability must be 1: he could not be in either half.

In Table 1.1, you can see a formula for Bayes' rule (there are different versions, but I shall just use this one) and how these numbers are plugged into it. The result is a posterior probability of .67. That is, Diesel's degree of belief is now lower than it was before they looked in through the windows, when it was at .8. That makes sense: Dick was *not* in the half they could see, which must reduce Diesel's confidence. However, there is still a greater than .5 probability that Dick is in there. The two sleuths now have a decision to make. We shall review the kinds of thinking that they might use to make it in Chapters 8 and 9.

In the rest of this chapter, we shall look at the research evidence on what people do when they think about probability, and at how this performance has been

Table 1.1 Inspector Diesel's belief revision, using Bayes' rule

The Bayesian formula:

$$\text{prob}(H|D) = \frac{\text{prob}(D|H) \times \text{prob}(H)}{[\text{prob}(D|H) \times \text{prob}(H)] + [\text{prob}(D|\neg H) \times \text{prob}(\neg H)]}$$

prob(H): Diesel's prior belief that Dick is in the Ham & Pickle 80% of the time, i.e. .8
prob(¬H): the prior probability of his alternative hypothesis that Dick is not there, i.e. .2
D: Dick is not in the visible half of the bar
prob(D|H): the likelihood that he is not visible, given Diesel's hypothesis, i.e. .5
prob(D|¬H): the likelihood that he is not visible, given the alternative hypothesis, i.e. 1

$$\text{prob}(H|D) = \frac{.5 \times .8}{[.5 \times .8] + [1 \times .2]} = \frac{.4}{.4 + .2} = .67$$

explained. We shall look firstly at 'plain' probability judgments, where people are asked to estimate likelihoods or frequencies, and then go on to belief revision, where we address the question of the degree to which people like you, me and Inspector Diesel are good Bayesians when we update our beliefs. So we shall be comparing normative (Bayesian) theory and descriptive accounts of human performance.

Judging plain probabilities

Dr Scale's first seven questions are about these, and you will be able to recall having thought about, or being asked about, similar probabilities many times. What are your chances of dying in an air crash compared to a car crash, or being struck by lightning? These questions are not always as easy to answer as they first appear, and explaining how we answer them is not always easy either.

Logical possibilities

People usually have a secure grasp of simple problems such as the ace of spades one above: in the classroom, I have not come across any people who have been truly baffled by questions like this. However, there is a sting in the tail even with these, and it is to do with people's understanding of randomness and what it implies for expected frequencies. People seem to assume that an equal chance means that there will be an even distribution, and are more surprised than they should be when the resulting pattern actually looks quite lumpy.

Here are two examples, one involving mass phenomena and one involving repetitive events. The first is quoted by Blastland and Dilnot (2007), in their splendid popular statistics book: cancer clusters. It is quite common to read in the press that there is anxiety about a higher than average incidence of cancer in some locality, sometimes following a change in the environment, such as the siting of a phone mast. With something as serious as cancer, people naturally want to know what

might have caused it. When there is no known cause, they will advance hypothetical causal models, such as possible carcinogenic properties of the radiation emitted by phone masts. (The World Health Organization, in 2009, advised that there is no convincing evidence that these induce or promote cancer, or any other illness.) They may even take action, such as sabotage, as Blastland and Dilnot report. But suppose cancer were distributed in the population at random, with respect to phone masts (we know quite a bit about actual causal mechanisms for some cancers, of course): what would the patterns of this distribution look like? They would not be even: that really would tell you that something fishy was going on. They would be lumpy: some areas would report a much higher incidence than average, some lower – they might even be right next door to each other. With real games of chance, such as coin tossing, people are uncomfortable with clusters of more than three, whereas 'lumps' of five or six heads or tails are quite common. Blastland and Dilnot report three sets of 30 coin tosses: there were three runs of five or six heads or tails in these sets.

The second example comes from the psychological literature, and was reported in a famous paper by Gilovich, Vallone and Tversky (1985), concerning people's beliefs about the 'hot hand' in US basketball players. This is a version of the general notion of 'form' in sports performers: that during a period of good form, a player is more likely to hit following a hit compared to following a miss than would be expected by chance. People strongly believe in form, and will swear that this elusive property is causing a higher than normal probability of hits. However, Gilovich and colleagues found no evidence for this effect in several studies of players' actual performance when they looked at the conditional probabilities of hits following hits and misses. What is happening, they say, is that people are oblivious of the statistical effect of *regression to the mean*. Each player will, over the long term, produce a personal average hit rate. In the short term, there will be deviations from this mean that are essentially random, and these deviations will sometimes occur in lumps. Just as with the cancer clusters, people then produce a causal hypothesis, form, to 'explain' the effect.

The *gambler's fallacy* is the most extreme version of this error. Think again about coin tosses, and suppose that there has been a run of five heads. It is quite common for people to believe that there is therefore a high probability of tails on the next throw, but, as the saying goes, the coin has no memory. The *logical possibility* of tails is still .5; this gives an *expected frequency* of 50/50, which is what it will approach in the long run. People who fall for the gambler's fallacy or the hot hand hypothesis are confusing the one for the other.

Alter and Oppenheimer (2006) review numerous studies of the hot hand fallacy and show how it and the gambler's fallacy can be seen as two sides of the same coin (forgive me). The hot hand idea incorporates the notion of skill as a causal mechanism, so that when there is a long streak of one particular outcome, such as hits, people expect the streak to continue. However, when the streak comes from a random, skill-free process such as coin tossing, people expect the streak to end, so that the general sequence balances out. An exception occurs with gamblers playing roulette or dice games in casinos: these are random processes and yet people do often believe in hot hands when gambling. This shows that gamblers have mythical beliefs about the processes that generate outcomes at the tables – a very dangerous state of affairs for the gambler, but a very happy one for the house.

Frequencies

We are now going to see how people approach Dr Scale's Questions 4, 6 and 7 and similar ones. One thing we can ask is how well calibrated people's estimates are compared to the statistical facts. We have already seen that there is a complication, which Question 4, about the weather forecast, brought out, and which applies when we try to use frequency expressions to convey single-event probabilities: you can be well calibrated in frequency terms but useless as a forecaster. Question 6, about robberies, is different. Now you are being asked a frequency question: how often do two things happen?

Evidence that people deviate from statistical norms comes from a classic piece of research by Lichtenstein, Slovic, Fischhoff, Layman and Combs (1978; see also Lichtenstein, Fischhoff & Phillips, 1982). They asked Americans in the 1970s to estimate how many deaths in the USA were due to various causes, such as different forms of disease, accidents, natural events (such as floods) and crime. People tended to overestimate the likelihoods of very uncommon events, such as deaths from botulism (a kind of food poisoning caused by infected meat), but underestimate the likelihoods of deaths from common causes, such as heart disease. As Baron (2008) points out, you would expect this kind of pattern anyway, given that people make mistakes: if something never actually happens, but some people think it does, its prevalence will come out as overestimated. The same will occur at the top end: anyone who mistakenly thinks that something does not happen, when it happens a lot, will pull down its average estimate.

There is a particular and very important difficulty with estimating frequencies: the *reference class* to which the event in question is being compared; this was mentioned earlier when discussing numerical weather forecasts. Remember that probability as frequency is reckoned by the number of times the event happens compared to the number of times it could have happened. This estimate clearly depends crucially on the latter number, so what should that number be?

Consider someone's chances of being struck by lightning. According to the BBC's weather website (accessed in October 2009), a Briton has a 1 in 3 million chance of being struck by lightning, and they make the point that this is a much higher chance than of winning the lottery jackpot (see above). However, it is hard to make sense of this claim without knowing what the 3 million refers to. This is the reference class problem. Probing further on the web, I found that about 50 people are struck by lightning in Britain every year (about 5 are killed). With a population of around 60 million, this is a strike frequency of about 1 in 1.2 million. But that is just in 1 year. In the USA, the 1-year risk is said to be 1 in 700,000, with a lifetime risk of 1 in 5000. However, even when we are clearer about such figures, we still do not have a clear reference class. These are the figures for national populations, and nations are made up of different kinds of people. Compare the chances of farm workers and miners: the former are obviously much more likely to get struck (one British agricultural worker has been struck seven times in 35 years). And if you live in Uganda, take especial care with your life insurance: Uganda has the most lightning storms of any country. So to judge your own chances of being struck, you need to know which class you belong to: there is a lot of difference between British

miners and Ugandan farmers in this respect. We shall come back to the reference class problem when we deal with belief updating.

Taking samples: where our information comes from

Think about Dr Scale's Questions 6 and 7 again, about street robberies and dogs and cats. People readily give responses to this kind of question, but what are they basing these responses on? There are two possibilities: prior knowledge and current information. We shall concentrate on the former, since that is where most of the research has been done. I doubt whether, when thinking about street robberies, you retrieved official government statistics, or, when thinking about dogs and cats, you had actually gone round your neighbourhood with a clipboard counting them up. So what did you use as the basis of your estimates?

The most influential answer to this question was proposed by Daniel Kahneman and Amos Tversky in an extensive research programme that has been compiled into two large volumes (Gilovich, Griffin & Kahneman, 2002; Kahneman, Slovic & Tversky, 1982). You use some kind of *heuristic*. A heuristic is a rough rule of thumb, as opposed to an algorithm, which is a set of exactly specified steps. If you are cooking, and faithfully following a recipe, you are using an algorithm; if you are vaguely tossing handfuls of this and that into the mixing bowl, you are using heuristics. The heuristic that applies to Dr Scale's two questions is *availability*, introduced by Tversky and Kahneman (1973). This is their own definition: 'A person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind' (p. 208). They add that you do not actually have to bring examples to mind, just estimate how easy it would be to do so. Thus one estimate, ease of thinking, is used to stand for another, probability.

Tversky and Kahneman argue that such a heuristic is ecologically valid (i.e. is true or useful in the real world), because if something is common it should be easy to recall instances of it. However, in some circumstances, its use can lead to biased estimates. This will happen if something other than frequency has led to availability. This is why Kahneman and Tversky's research has come to be known as the *Heuristics and Biases* programme. It is as if the mind commits a logical fallacy (we shall deal with these in Chapter 3): given that *If something happens often then examples will be easy to recall*, and *I can easily recall an example*, you infer *therefore it is happening often*. This would only be valid if nothing else led to ease of recall. The observation of such biases would be crucial evidence that people use heuristics when judging probability.

This prediction was confirmed in a series of experiments, the hallmark of which is their ingenuity and simplicity (see Part IV of the Kahneman et al., 1982 volume and Part One of the Gilovich et al., 2002 volume for collections of these studies). Here is an example from Tversky and Kahneman's original 1973 paper. They chose the letters, K, L, N, R and V, and asked 152 people to judge whether each was more likely to appear as the first letter or as the third letter in a sample of English text. More than two-thirds (105) thought that the first position was more likely for a majority of the letters. In fact, all the letters are more common in the

third position. The availability heuristic would predict this since we are far more used to thinking about initial letters, such as in alphabetical name lists, than third letters. In the same paper, they gave people two lists of names. In one there were 19 women and 20 men, and in the other, 19 men and 20 women. In both cases, the 19 names were famous and the 20 were not. When asked to estimate which were more frequent in the lists, men or women, the participants reported that there were more women in the first list and more men in the second – the opposite of the truth. They were biased by the greater ease of recalling the famous names, as was confirmed in a memory test when they recalled 50% more of these than the unfamiliar names.

Evidence that availability can affect more real-life judgments comes from a study by Ross and Sicoly (1979). It concerned people's perceptions of their share of the credit when they have joint responsibility for some event or activity: students will think about this when doing group projects, for instance, and sports team members will ponder who contributed most to their team's performance. Ross and Sicoly studied married couples, thinking about 20 household events (19 for those without children): who did each partner think did the most of each? He and she both thought they made a greater contribution to most of the events, including negative ones such as starting arguments. They cannot both be right. Ross and Sicoly ruled out a motivational bias, because the effect occurred with both good and bad events. The availability explanation is that your own contributions are simply easier to remember: you are always there when you do something, but not when your partner does, and you may also attach greater significance to your own deeds.

Anything that makes an event easy to bring to mind should increase our judgment of its probability whether or not it has anything to do with frequency, according to availability. An example is *vividness*: how much impact does the information make on you? Nisbett and Ross (1980) devote a whole chapter of their classic book to this factor. The street robbery question above was designed to illustrate it. If you thought that old ladies are more often victims than young men are, you may well have thought so because you could easily call to mind vivid instances: you may have seen news stories with graphic pictures, and these are not easy to forget. The memorability of such stories may in turn be due to a factor that Nisbett and Ross identify as contributing to vividness: emotional impact. Proximity in time or space is another: for instance, if the houses on either side of yours have noisy dogs in them.

They also point out that vivid information is usually more information, for instance in the reporting of an elderly crime victim, where a lot about the person and the circumstances of the crime will be given. Celebrity illness is another instance of this: it was observed by Nisbett and Ross over 30 years ago, and can be observed today, that a famous person reported as suffering from a particular disease leads to a flood of people seeking tests for that disease. Politicians and commentators know all about vividness: you will often hear them invoking single striking cases to attack an opposing view, when the statistics actually back up that view. Nisbett and Ross sum up the vividness factor with a chilling quote attributed to the Soviet dictator Josef Stalin: that a single death is a tragedy, while a million deaths is a statistic.

There is another way in which factors such as vividness could affect probability judgment without the availability heuristic: they could bias the sample on which you base your judgment. Think of the news reports of elderly crime victims:

how often do such stories appear, compared to stories about young male victims? It is possible that, when making the kinds of judgments that Question 6 called for, you are responding accurately to the sample that the media present you with. News stories are often about unusual occurrences – that is what makes them news – and so the unusual will come to be represented more often in current experience and memory than it should be, statistically speaking. This would explain the apparent bias in people's assessments of likelihood of death from various causes: they will overestimate rare causes because they are talked about more than common causes, and underestimate common causes for the same reason. Similarly, consider Ross and Sicoly's married couples. Perhaps each thinks they do more than the other simply because they observe more of their own contributions than their partners'.

Klaus Fiedler (2000), in an extensive survey, makes the point that our judgments are always based on samples, and that if the sample that is presented to us is biased, as with sexy stories in the media, then so will our judgments be, especially if our judgment processes themselves tend to be accurate. The bias is in the sample, not in the mind. What we need in order to evade bias is a skill that seems very hard to acquire: that of critically assessing the nature of sampling. The letter task (K, L, N etc.) used by Tversky and Kahneman (1973) in their initial proposal of availability (see above) was criticised on sampling grounds by Sedlmeier, Hertwig and Gigerenzer (1998). They point out that the set of letters used in this study was a biased sample of letters of the alphabet: all five are more common in the third position in English words, whereas 60% of all consonants are in fact more common in the first position. Sedlmeier et al. proposed that people would be sensitive to the relative frequencies within the whole class of consonants, while tending to underestimate the most common letters and overestimate the rare ones, as in the death estimates in the Lichtenstein et al. (1978) study reviewed earlier. A model based on this hypothesis predicted people's judged ordering of frequencies of 13 letters in German words better than did two versions of the availability heuristic. We shall come back to the question of sampling later in the chapter.

The notion of availability was elaborated in later work by Tversky and his colleagues, in the form of *support theory*. Tversky and Koehler (1994; see also Rottenstreich & Tversky, 1997) distinguished between the event itself and mental representations (or descriptions, of events) which they called *hypotheses*. When you judge a hypothesis, you consider the weight of evidence for it: its support. Hypotheses can be explicit or implicit. For instance, suppose you were asked to judge how many people die each year of natural versus unnatural causes. You are given heart disease, cancer and other natural causes to assess; or accident, homicide and other unnatural causes. In each case, the named causes are explicit hypotheses and 'other' is an implicit hypothesis. Explicit mention of factors such as accident or homicide provides cues to search for support that are absent from the implicit hypothesis: it is hard to search for 'other'.

Both earlier research and experiments run by Tversky and Koehler (1994) showed that when implicit hypotheses are unpacked into explicit components and their probabilities judged, the sum of judgments of the explicit hypotheses is greater than the judgment of the implicit hypothesis. This is called *subadditivity*: the implicit support comes to less than the quanta of explicit support when you add them up. It follows from this that the degree of subadditivity should be greater

when there are more unpacked explicit components. Objectively, of course, there should be no difference: 'natural causes' should just be the sum of all the causes that come into this category. Both tendencies, subadditivity and its increase, were confirmed by Tversky and Koehler (1994). In one experiment, they asked US students to estimate the probability of death from various causes, either by thinking about the probability of an individual dying from them, or by assessing the frequency of each from the 2 million annual deaths in the USA. These students assessed the probability of 'natural causes' as .58, but their estimates of 'heart disease + cancer + other natural causes' came to .73, a subadditivity factor of 1.26 (i.e. .73/.58). When seven rather than three components were used, subadditivity was increased, as support theory predicted: the factor was now 2.19. The same tendencies were found for unnatural causes.

Interestingly, and consistent again with the findings of Lichtenstein and colleagues (1978) mentioned earlier, the unnatural causes were greatly overestimated relative to the actual frequencies of recorded causes of death. For instance, estimates of frequency of deaths due to accidents were around 30%, but the true figure was 4.4%. Support theory is not about people's accuracy relative to the facts, but about the way they judge their *representations* of events, their hypotheses. These hypotheses can only be formed about what is available, hence unpacking and subadditivity can be affected by the sorts of factors, such as vividness, that influence availability.

Belief updating

In many everyday cases of probability judgment, we do not judge probabilities by themselves, but judge them in response to some information. That is, we start out with some estimate of belief and then have to revise it. Medicine is an obvious case: you have some symptom that leads you to suspect that you may have a certain disease, you get yourself tested, and the test result comes back. Dr Scale has one last problem for you to illustrate this, adapted from an example in Eddy (1982):

9 A friend of yours went on holiday to the Costa del Sol last year and, in between bouts of inadvisable partying, fried in the sun on the beach for two weeks. Recently, he noticed a large new mole on his arm. He went to the doctor, who decided to test him for skin cancer. She told him that (a) in people who have cancer, the test shows positive in 90% of cases, while (b) the false positive rate, where people without cancer test positive, is 20%. People with new moles like your friend's actually turn out to have cancer 1% of the time (c). Your friend is shaken: his test has come out positive. He wants you to tell him what the probability is that he has cancer. What is your answer?

Give a quick estimate before going on: that is what your friend did. And then work it out. As with the Diesel and Roscoe problem (Question 8) earlier, you have all the information in front of you to work out a precise numerical answer, using Bayes' rule, which was given in Table 1.1. To help you along, the prior probability that he has cancer, prob(H), is .01 (c above). The probability that he will test positive if he

Table 1.2 Bayesian answer to the skin cancer problem

The test shows positive in 90% of cases in patients who have cancer: prob(D|H) = .9
The test shows positive in 20% of cases in patients who do not have cancer: prob(D|¬H) = .2
1% of people like this patient actually have cancer: prob(H) = .01

Using Bayes' rule from Table 1.1:

$$\text{prob}(H|D) = \frac{.9\times.01}{[.9\times.01]+[.2\times.99]} = \frac{.009}{.009+.198} = \frac{.009}{.207} = .043$$

has cancer, prob(D|H), is .9 (a), while the probability that he will test positive even though he is clear, prob(D|¬H), is .2 (b). You can now put these numbers into the Bayesian formula in Table 1.1 and work out the posterior probability, prob(H|D), that he has cancer given a positive test result. It will do you good if you do, but Table 1.2 shows the workings if you want to check or avoid it.

The answer is that prob(H|D) = .043, or just over 4%. Or to put it another way, the probability that he does not have cancer is .957: he almost certainly does not have it, even though he tested positive. The test therefore is useless.

You probably find that result surprising, and your friend may take some convincing of it too. This intuition is a clue as to why there has been a huge volume of research into belief updating. If your initial estimate was around .9, which is the result many people give, you have committed the *inverse fallacy* (Koehler, 1996): you have given prob(D|H) instead of prob(H|D). There are two reasons why the latter figure is so low: the false positive rate of 20% is high, and the prior probability of .01 is low. The latter is known as a *base rate*: the amount of the disease that there is in this population to begin with. It looks as if people tend to neglect base rates with problems like this. Base rate neglect is the most researched aspect of belief updating, so we shall begin our detailed examination of this kind of judgment with it. If you found the Bayesian descriptions above rather hard going, rest easy: later on you will see a method that makes deriving the answer to problems like this much clearer – not just for people like you and me, but for experts such as doctors too. The method in question is at the very heart of debates about just how people revise their beliefs, and what determines their accuracy when they do so.

Base rates: neglect or respect?

The medical problem just given shows that it is possible to induce people to neglect base rates when updating belief in the light of evidence. Eddy (1982) actually tested doctors with this kind of medical problem. You might imagine they would be less prone to base rate neglect than non-medics in this context. They were not. This rather alarming state of affairs shows how potentially important base rate neglect is.

As with so much research in this area, the modern study of base rate neglect was kick-started by Kahneman and Tversky in the 1970s. Neither they nor Eddy invented the diagnosis problem (credit for that goes back to Meehl & Rosen, 1955),

but characteristically they did introduce some striking new experiments and, equally importantly, a provocative theory to account for their findings. Both their methods and their theory have generated hundreds of research papers and volumes of debate, culminating in two major reviews in the journal *Behavioral and Brain Sciences*, by Koehler (1996) and Barbey and Sloman (2007); each paper is followed by dozens of commentaries from other experts.

Kahneman and Tversky (1973) introduced one kind of problem that has formed part of the bedrock of research into base rate neglect: the personality problem. It comes in two guises. In the first, participants were told about a room full of 100 people, some of whom are engineers and some lawyers. They were given a personality description of an individual said to have been picked at random from the hundred, and asked to judge whether it is more likely that he is an engineer or a lawyer; the description was designed to resemble the stereotype of engineers, say. One group of participants was told that there were 70 lawyers and 30 engineers in the room, while another group had the figures reversed, so that there were 30 lawyers and 70 engineers. These were the base rates, and they should have influenced judgments: the random person is more likely to be an engineer than a lawyer in the second group than the first. But they made no difference: each group's judgment of how likely the random man was to be an engineer was the same.

In the second version, 69 students were asked to estimate the percentage of students across the country studying nine academic subjects. This provided the base rates, as they were understood by these participants. A second group drawn from the same student population was given a personality description of 'Tom W.', portraying him as something of a nerd, and asked how similar he was to the typical student in each of the nine fields. Finally, a third group was told about Tom and asked to predict which field of study he was likely to be involved in. Their predictions correlated almost perfectly with the second group's similarity judgments, and were negatively related to the first group's base rate judgments. For instance, 95% of the prediction group judged that Tom was more likely to be studying geeky computer science than Bohemian humanities and education, but the base rate of the latter was three times that of the former.

Faced with these results (and others), Kahneman and Tversky proposed that the people were using a non-statistical heuristic to arrive at their judgments: *representativeness*. This word is related to ideas such as similarity and resemblance. In an earlier paper, they defined representativeness in this way:

> A person who follows this heuristic evaluates the probability of an uncertain event, or a sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated.
>
> (Kahneman & Tversky, 1972, p. 431).

The personality tasks are to do with (i): people are noting the similarity between the individuals described and the typical features of a stereotyped engineer or computing student. These are the parent populations, as far as the participants are concerned.

The second kind of representativeness was explored using perhaps the best known base rate task of all, the taxicab problem. Here it is:

> A taxi is involved in a hit-and-run accident at night. In the city, there are two taxi firms, the Green Cab Company and the Blue Cab Company. Of the taxis in the city 85% are Green and the rest are Blue.
>
> A witness identifies the offending cab as Blue. In tests under similar conditions to those on the night of the accident, this witness correctly identified each of the two colours 80% of the time, and was wrong 20% of the time.
>
> What is the probability that the taxi involved in the accident was in fact Blue?

We pause here for you to work out the Bayesian answer. If you need some help, look at Table 1.3, where, as with the Diesel and Roscoe and diagnosis problems, the numbers just given are put into the Bayesian formula.

We are trying to work out the probability that the taxi was Blue given that the witness said it was Blue: prob(H|D). For that, we need the prior probability, prob(H), that the cab was Blue. From the problem description we can infer this to be .15, the figure for 'the rest' once the 85% Green taxis are taken into account. Now we need the data, D. This is the witness's testimony, and he or she was 80% accurate. So we compare the times when the witness says Blue and it really is Blue to the times when he or she says Blue and it is actually Green, taking into account the proportions of Blue and Green cabs.

In Table 1.3 you can see these various bits of information put into the Bayesian formula. The result is that the posterior probability that the cab was in fact Blue is .41 which means that, on these figures, the cab was actually more likely to be Green (.59)!

Tversky and Kahneman (1982a) report data from experiments in which people were given the taxicab problem. They report that, like the personality problems, judgments were largely unaffected by base rate: most were around .8, which is the figure for the witness's accuracy, and is prob(D|H). This is a case of what

Table 1.3 Bayesian answer to the taxicab problem

What is the probability that a taxi involved in an accident was Blue, given that a witness identified it as Blue: prob(H|D)?

The city's cabs are 85% Green and 15% Blue: prob(H) = .15, prob(¬H) = .85
The witness is accurate 80% of the time and mistaken 20% of the time
prob((D|H), that the witness says Blue when the cab is Blue, is therefore .8
prob(D|¬H), that the witness says Blue when the cab is Green, is .2

Using Bayes' rule:

$$\text{prob}(H|D) = \frac{.8 \times .15}{[.8 \times .15] + [.2 \times .85]} = \frac{.12}{.12 + .17} = \frac{.12}{.29} = .41$$

Kahneman and Frederick (2002) call *attribute substitution*, as is the use of the availability heuristic: we answer a question to which the answer is inaccessible by addressing another one, whose answer is more accessible. In the availability problems we think about recallability: in the cab problem we think about the witness's reliability.

Base rate neglect is in fact only one facet of the inverse fallacy. With base rate neglect, people are underweighting prob(H), but, as Villejoubert and Mandel (2002) point out, the inverse fallacy also involves neglect of prob(D | ¬H) – the false positive rate in the diagnosis problem, or the witness's mistaking Green cabs for Blue in the taxicab problem. In an experiment, they showed that judgments varied according to the degree of difference between prob(H | D) and prob(D | H) with base rates held constant, thus confirming that the inverse fallacy and base rate neglect can be teased apart. We shall return to other aspects of the inverse fallacy shortly, when we consider a radical alternative proposal to Kahneman and Tversky's heuristics in explaining belief revision. We shall turn from medicine to the law when we do so.

Belief revision by natural frequencies

Are you still worried about those doctors that Eddy (1982) tested, who were just as bad at judging the probability that a patient has a disease, given a positive test result, as you were? Relief is at hand. Eddy gave his medical participants Bayesian problems described in the way they have been here: in words. Think about the relation between these word problems and probability judgments in real life. Two differences jump out straight away: the words themselves – is this the way real people talk about probability?; and the data the problems present – in real life, we don't often encounter summary statistics, but build up our representation of the statistical world bit by bit. This is called *natural sampling*. Might doctors do this? Christensen-Szalanski and Beach (1982) found evidence that they do: doctors who learned the relation between the base rates of disease and the outcomes of tests through their clinical experience did not neglect base rates when making diagnostic judgments.

Natural sampling is at the core of a gigantic research programme into probability judgment, decision making and rationality conducted since the 1980s by Gerd Gigerenzer and his colleagues. As this list of areas implies, we shall look at this research not only in this chapter but in later chapters too, especially Chapters 9 and 10. For the moment, let us consider the Gigerenzer approach to probability judgment and belief revision.

Gigerenzer is a frequentist, and frequentists accept only repeated observations as the basis for probability judgments; they do not accept that there is any coherent way to judge the probability of unique events. In his landmark paper with Ulrich Hoffrage, Gigerenzer gives an evolutionary justification for this position:

> Evolutionary theory asserts that the design of the mind and its environment evolve in tandem. Assume. . . that humans have evolved cognitive algorithms that can perform statistical inferences. . . For what information format were these algorithms designed? We assume that as humans evolved, the 'natural'

> format was *frequencies* as actually experienced in a series of events, rather than probabilities or percentages.
>
> (Gigerenzer & Hoffrage, 1995, p. 686)

The point at the end is that probabilities (from 0 to 1) and percentages are recent cultural developments, whereas natural frequencies could, they propose, have been used by any person at any time in history. It is important to be clear about natural frequencies. It is not that any frequency format will make Bayesian belief revision easier, but that *natural* frequencies will. For instance, relative frequencies will not have the effect, because these are 'normalised' numbers, not instances actually encountered. Percentages, such as those you have just seen in the diagnosis and taxicab problems, are a form of relative frequency: 85 out of every 100 cabs are Green, and so on. This is not the same as saying that 100 cabs have been observed, of which 85 were Green. The dice-playing nobleman we heard about at the beginning of this chapter was using natural frequencies.

Now let us re-word the diagnosis problem set out above in terms of natural frequencies. Take 1000 patients who see their doctor about a new mole on their skin. Of these people, 10 (the 1% given in the original problem) actually have skin cancer. The test will give a positive result for nine of these (90% of them). There will be 990 people who do not have cancer, and the test will show positive in 198 cases (the 20% false positive rate). So what is the likelihood of cancer when given a positive test? There are 207 people with positive test results (9 + 198), of whom 9 have cancer; 9 out of 207 = .043.

That wasn't so hard, was it? If you look back at Table 1.2 you will see these figures, in decimal form, emerging from the Bayesian formula. It is not the figures that are different, but their representation. Gigerenzer and Hoffrage make this representation even clearer by using a tree diagram, and in Figure 1.1 the diagnosis and the taxicab problems are represented in this way. As you can see, and as they emphasise, the computations needed to derive the Bayesian answer with the natural frequency format are very much easier than they are with Bayes' rule itself. In fact, they reduce to

$$\text{prob}(H \mid D) = \frac{a}{a+b}$$

where a is the frequency of prob(D|H) observations – people with positive test results who actually have cancer, cabs identified as Blue which really are Blue – and b is the frequency of prob(D|¬H) observations – positives without cancer, wrongly identified cabs. Notice what is missing here: prob(H), the base rate. If you represent these problems in natural frequency terms, *you don't need a separate expression for the base rate*; since natural frequencies are derived from the base rate, they contain base rate information within themselves (see Kleiter, 1994, for earlier steps in this argument).

In an experiment, Gigerenzer and Hoffrage (1995) confirmed that presenting problems like this (they used 15 problems, including the cab problem and various medical ones) in a frequency format that promoted representation of just the a and

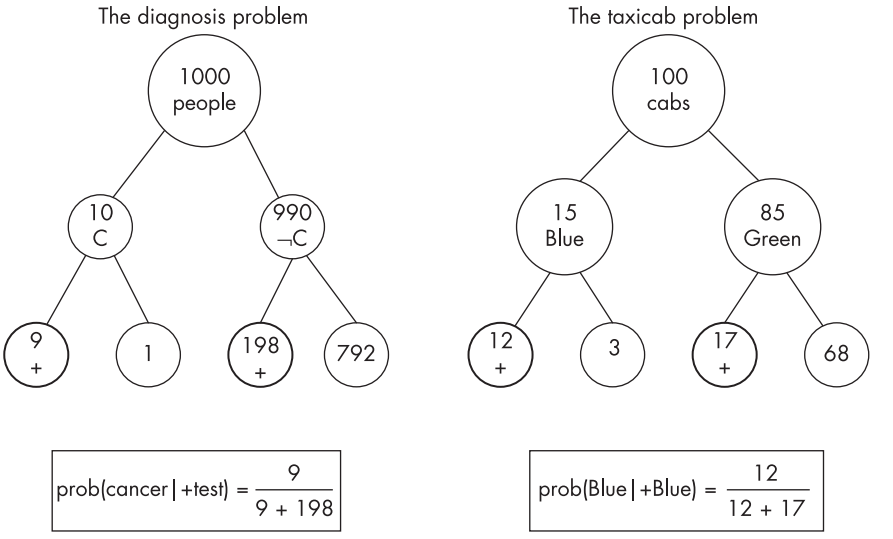


Figure 1.1 The diagnosis and taxicab problem in natural frequency form

Note: In the diagnosis problem + means prositive test, C means has cancer and ¬C means does not have cancer; in the taxicab problem, + means witness says Blue

Source: Adapted from Gigerenzer and Hoffrage (1995)

b components resulted in significantly more Bayesian answers than did the standard probability format: 50% versus 16%.

Bear in mind that the question their participants were asked was not the likelihood that a single person had a disease or that a single accident involved a Blue taxi: that would be asking for a single-event probability. In order to ask this, you really do need the base rate, because base rates depend on reference classes (see above): a base rate frequency is the proportion of times an event occurs out of the times it could have occurred. The latter part depends on the reference class. The trouble with this is that there are infinitely many possible reference classes for any single base rate. Consider the diagnosis problem. What is the right reference class for your friend with the mole? Men, young men, young men who have recently been sunburned, young sunburned men with a particular skin type? It is impossible to say. Ultimately, your friend is his own reference class. In natural frequency experiments, people are asked how many ___ out of ___ have the disease, were Blue, and so on. The blanks are filled in with absolute frequencies (i.e. simple counts of actual occurrences), which do not depend on reference classes.

Gigerenzer is not the only researcher to address belief revision in this way. Two leading evolutionary psychologists, Leda Cosmides and John Tooby (1996), also conducted a series of experiments on frequency formats. They used variations on the diagnosis problem. In their basic reformulation of the problem, they found that 56% of people produced the Bayesian solution with a frequency format, a figure comparable to Gigerenzer and Hoffrage's result. However, they included

redundant percentages in their experiment (e.g. including the phrase 'a "false positive" rate of 5%' alongside the frequency of 50 out of 1000 people testing positive without the disease) and discovered that performance rose to 72% when this information was removed. Using further manipulations, such as presenting the problem information pictorially as well as in figures, they raised the rate of Bayesian responding as high as 92%. We shall return to this particular factor below.

Not surprisingly, results like these, along with the theory that predicted them, have generated a large research literature. It is the latter, the theory, that explains this boom. Frequency formats had appeared in research papers before, sometimes to explain the problems to readers (Hammerton, 1973; Pollard & Evans, 1983), sometimes as part of the experimental materials (Fiedler, 1988; Tversky & Kahneman, 1983), but Gigerenzer's approach offered a rationale for their success – a rationale that has not gone unchallenged, as we shall see.

Gigerenzer and his colleagues have devoted enormous energy to their research programme, and I shall give some more examples of this work here. It has been collected into a series of books, for specialist and non-specialist audiences. For specialist references, go to Gigerenzer and Todd (1999a) and Gigerenzer and Selten (2001); for more 'popular' treatments, try Gigerenzer (2002, 2007).

This work has obvious potential applications, because people need to make probability judgments in order to make informed decisions. We shall look closely at decision making in later chapters, and will return to Gigerenzer's approach to it when we do so, but for now let us look at two areas in which the natural frequency approach has been applied in the real world: health and the law. In both cases, it is no exaggeration to say that probability judgment can be a matter of life and death.

In his popular book, Gigerenzer (2002) recounts a personal experience. He needed to take an HIV test as a condition of obtaining a work permit for the USA. HIV, the virus that leads to AIDS, is rather like lightning: your chances of being struck by it depend on who you are and what you do. So the first thing to work out is your reference class. Gigerenzer's was low-risk German males, to whom the following data applied at the time:

Base rate: .01%
Test sensitivity: 99.9% (i.e. chance of testing positive if you have HIV)
False positive rate: .01% (i.e. chance of testing positive even though you do not have HIV)

Now suppose such a person tests positive: what is the chance that he has HIV? You could construct a tree diagram like the one in Figure 1.1, but you hardly need to in this case. You can use natural frequencies in your head. Start by thinking of 10,000 low-risk German men. How many have HIV? In frequency terms, .01% is 1 in 10,000, so the answer is 1. He will be detected with almost total accuracy (99.9%, or .999). Now think about the remaining 9999 men. The risk of testing positive is also 1: 10,000, so we can be almost certain that there will be one misfortunate man among them. How many positive test results do we have then? Two, of which one is true. On these (real) figures, his chances of actually having HIV, given a positive test result, would be 1:2. He should take another test.

The really important figure in making these judgments is the false positive rate. These can only be derived from observed frequencies because they are by nature unpredictable. We then need to know how to use them in our calculations. As just mentioned, these judgments can have the most profound consequences for people: there is a big difference between thinking there is just a 50/50 chance of having HIV and thinking it is almost certain that you have it. You might not get your work permit; you might, if you are sure you have HIV after testing positive, wind up a victim of suicide, assault or even murder (such incidents are related in Gigerenzer, 2002).

One thing you can be sure of when taking an HIV test is that you will want some good advice. Gigerenzer, Hoffrage and Ebert (1998) conducted a field study of AIDS counsellors in Germany: Ebert had himself tested and counselled at 20 health centres, presenting himself as a low-risk client. You would hope, indeed expect, that counsellors would know how to interpret test results. The ones in this study did not. Ebert asked the counsellors about a number of aspects of the test, such as its sensitivity and false positive rate, and, most importantly, what the result would mean for the likelihood that he actually had HIV, if it came out positive. Five of the counsellors said that the test was 100% sensitive, while thirteen were accurate on this point; only three were clear about false positives from the beginning; while three-quarters told Ebert that he was certain or almost certain to have HIV if he tested positive (against the true figure of 50%). Needless to say, Gigerenzer et al. recommend that counsellors be retrained to interpret, and talk to their clients about, natural frequencies.

Now for the law. In 1999, Sally Clark, an English lawyer, was convicted of the murder of two of her sons. Each had been found dead in their cots when only a few weeks old. Part of the case against her was the testimony of an eminent paediatrician, who told the court that the probability of two children in the same middle-class family dying of sudden infant death syndrome, or SIDS (a residual category, not a diagnosis, arrived at after all other possible causes such as injury and disease have been eliminated), was 1:73 million. Clark served over 3 years in prison before her conviction was quashed on a second appeal in 2003. At that appeal, two sorts of statistical error were pointed out. Firstly, the 73 million: this was arrived at by multiplying the single probability of one SIDS death, about 1:8500, by itself, just as we did with the calculation of 'snake eyes' at the start of this chapter. This is a factual mistake: SIDS deaths are not random events. There are factors within families that make some more prone to suffer this tragedy than others, so if there has been one SIDS death the probability of another is much higher than the random probability. Secondly, the jury might well have committed the *prosecutor's fallacy*: using the paediatrician's figure as the probability that Clark was innocent (see Nobles & Schiff, 2005, for a brief account of this case).

The prosecutor's fallacy is a version of the inverse fallacy: mistaking prob(D|H) for prob(H|D). Even if the figure given was correct, it is the expression of the probability that two children will die in this way (D) given that the alternative in this case – murder – has been eliminated (i.e. that the defendant is innocent (H)). It is *not* the probability that the defendant is innocent given these deaths. Jurors should have compared any such estimate with the prior probability of the alternative hypothesis: that two babies were murdered by their mother. This is

likely to be even lower. Here is a parallel, again reaching back to the start of this chapter. You have won the lottery, the odds against this happening being 1:13.98 million. Someone accuses you of cheating. The odds just quoted are the odds that you will win (D) given that you don't cheat (H); they are not the odds that you didn't cheat given that you won. We need to know the prior probability that you cheated, and this is vanishingly remote, because of the design of the lottery machine. Jurors could no doubt see this quite readily, and yet they have fallen prey to the inverse fallacy in cases like Clark's (other women have had similar convictions overturned since her case) – Sally Clark died of alcoholic poisoning 4 years after her release from prison.

The inverse fallacy bedevils the use in criminal trials of the most important aid to crime investigation to have emerged in a century: DNA profiling. TV dramas present this as a failsafe index of a suspect's culpability, but there is no such thing as an infallible test. Even if there is a tiny chance of error, as in the case of the HIV statistics above, this must be taken into account. And these chances must be presented to juries in a way that their minds can handle, otherwise miscarriages of justice will happen.

Suppose that a sample of DNA is taken from a crime scene, it matches the suspect's DNA profile, and the probability that a person selected at random would also match the sample (the random match probability) is 1:1 million. Does that mean that the odds against the suspect's innocence are a million to one? No. Once again, 1:1 million is prob(D|H), the probability that the sample would be found if the suspect were innocent, not prob(H|D), the probability that he is innocent given that this sample has been found. We need to know some other things in order to compute prob(H|D): we need the false positive rate, or prob(D|¬H), and the prior probability that the suspect could have been the perpetrator in the first place, prob(H). A cast-iron alibi, for instance, drastically reduces this figure. And we need to know about the prior probability of the alternative hypothesis, that someone else did it.

Lindsey, Hertwig and Gigerenzer (2003) report a study in which over 100 advanced law students and academics were presented with trial statistics, based on real cases, in frequency and probability formats. The two versions are given in Table 1.4. The first two questions concern the interpretation of the statistical information. Question 1 requires thinking about false positives, and with the probability version performance was appalling: hardly any of the students or academics found the correct answer, which is .09 (there are 110 men with a reported match, of whom 10 actually have a matched profile; 10/110 = .09). These figures rose to 40% (students) and over 70% (academics) with the frequency presentation. The picture was almost identical with Question 2: while again the academics were better than the students, the difference between the probability and frequency conditions for both groups was massive (the probability is .0091, or 1 in 110). And the verdicts (Question 3)? Fifty per cent more law students and academics returned a guilty verdict in the probability condition. Their level of reasonable doubt was higher in the frequency condition. Now, imagine you are the suspect in the previous paragraph, who happens to have a DNA profile that matches the one found at the scene; there is no other evidence against you; you are, in fact, innocent. You live in or near London, where there are about 10 million people. This means that there will be

Table 1.4 Probability and frequency versions of the DNA problem in Lindsey et al. (2003)

Probability version

The expert witness testifies that there are about 10 million men who could have been the perpetrator. The probability of a randomly selected man having a DNA profile that is identical with the trace recovered from the crime scene is approximately 0.0001%. If a man has this DNA profile, it is practically certain that a DNA analysis shows a match. If a man does not have this DNA profile, current DNA technology leads to a reported match with a probability of only 0.001%.

A match between the DNA of the defendant and the traces on the victim has been reported.

Question 1. What is the probability that the reported match is a true match, that is, that the person actually has this DNA profile?
Question 2. What is the probability that the person is the source of the trace?
Question 3. Please render your verdict for this case: guilty or not guilty?

Frequency version

The expert witness testifies that there are about 10 million men who could have been the perpetrator. Approximately 10 of these men have a DNA profile that is identical with the trace recovered from the crime scene. If a man has this DNA profile, it is practically certain that a DNA analysis shows a match. Among the men who do not have this DNA profile, current DNA technology leads to a reported match in only 100 cases out of 10 million.

A match between the DNA of the defendant and the traces on the victim has been reported.

Question 1. How many of the men with a reported match actually do have a true match, that is, that the person actually has this DNA profile? [sic]
Question 2. How many men with a reported match are actually the source of the trace?
Question 3. Please render your verdict for this case: guilty or not guilty?

10 people from this region who match: you and nine others. The odds on your innocence are 9:1, not 1:1 million. If you ever find yourself in the dock, you had better hope your lawyer knows how to talk to juries about probability.

People tend to find that the re-presentation of probability problems in natural frequency formats makes them strikingly clearer, so you may be wondering whether there have been any moves to incorporate this way of dealing with probability into the education system. As Bond (2009) reports, there have: Gigerenzer himself has been involved in educational programmes with doctors and judges, and primary school children in Germany have been given classes where they manipulate frequency information. However, we should remember one of the messages of the previous few pages: consider alternative hypotheses. Sure, natural frequencies can lead to clearer probability judgments, but how? Is it possible that they have their effect through a different mechanism from the one favoured by Gigerenzer's school?

Probability from the inside and the outside

Not all psychologists in this area have accepted the frequency research and theory – particularly the theory, with its argument for an evolutionary adaptation for understanding frequencies but not single-event probabilities. The natural frequency theory would be in trouble if it could be shown either that this format did not facilitate probability judgment or that some probability formats did. Evans, Handley, Perham, Over and Thompson (2000) claimed to have found little difference between hard and easy frequency formats and a probability format. However, their 'hard' frequency format presented normalised frequencies (such as the 85 in every 100 referred to above), which Gigerenzer and Hoffrage have always maintained would not facilitate because they obliterate base rate information. The low level of performance with the 'easy' problem, at 35%, is not easy to explain. Evans et al. used tasks based closely on those used by Cosmides and Tooby (1996), which, as we saw earlier, sometimes produced different levels of performance from those observed by Gigerenzer and Hoffrage. Sloman and Over (2003) also report not obtaining such high levels of performance as Gigerenzer and Hoffrage and Cosmides and Tooby did. This may be the most important point from these various studies: natural frequency formats do not always bring forth accuracy from a majority of the people presented with them, which is a problem for a theory that says they should.

Girotto and Gonzalez (2001) tested natural frequency presentations against a novel probability format: chances. It is novel for psychological research, but, as Girotto and Gonzalez remark, this format is often used in real life, and it was used in considering Dr Scale's card-playing questions earlier: saying that there are 4 chances in 52 of finding an ace is expressing probability in this way. They found that it promoted accurate judgments as well as natural frequencies did. Hoffrage, Gigerenzer, Krauss and Martignon (2002) counter that the chances format merely mimics natural frequencies, something that Girotto and Gonzalez do not accept, using playing cards as an example: such problems are easy, but they are not easy because you have sampled sets of cards, but because you know about their logical proportions in advance.

What all these authors focus on, both the Gigerenzer school and its critics, is an idea originally put forward, yet again, by Kahneman and Tversky: that probability problems asking for judgments of single events – this patient, this taxi – encourage a focus on the properties of the individual case in question, an 'inside' view, while frequency problems encourage a focus on the class of which the individual case is a member, an 'outside' view (Kahneman & Tversky, 1982b). We shall see how this applies to base rate problems, with which we have been mainly concerned up to now, and then look at some others that have figured prominently in the literature: the planning fallacy, overconfidence, and the conjunction fallacy.

As far as base rate problems are concerned, both Evans et al. and Girotto and Gonzalez strongly emphasise what has come to be known as *nested-sets theory*. This has also been advocated in the recent review by Barbey and Sloman (2007). Gigerenzer and his colleagues (see Gigerenzer & Hoffrage, 2007; Hoffrage et al.,

2002) argue vehemently that this has always been an inherent aspect of the natural frequency theory in any case. It can be stated very easily. Refer back to the little equation above, which expresses what you need to compute from the information provided in Gigerenzer and Hoffrage's tree diagrams (see Figure 1.1): prob(H|D) = $a/a+b$. What Gigerenzer's critics argue is that anything that helps you to see that a is part of $a+b$ should help you to the Bayesian solution, and that natural frequencies just happen to do this, whereas Gigerenzer argues that natural sampling is the only reliable way to this insight. The diagrams used by Cosmides and Tooby (see above) may have helped in this, and in producing such high levels of performance.

Sloman and Over (2003; see also Sloman, Over & Slovak, 2003) also report that diagrams aid Bayesian reasoning, but only when attached to tasks presented in single-probability format: they did not boost the already good performance brought about by frequency presentation, unlike in Cosmides and Tooby's research. Importantly, they also used purely verbal means to clarify the nested-sets relation, for example:

> The probability that an average American has disease X is 1/1000. A test has been developed to detect if that person has disease X. If the test is given and the person has the disease, the test comes out positive. But the test can come out positive even if the person is completely healthy. Specifically, the chance is 50/1000 that someone who is perfectly healthy would test positive for the disease.
>
> Consider an average American: what is the probability that if this person is tested and found to have a positive result, the person would actually have the disease?

The answer is 1/51, or 1.8%, or .018. Forty per cent of participants produced an answer at or near this value, a figure close to that commonly observed with frequency formats. Notice that the information is given in the form of normalised relative frequencies, and the question asks for a single-probability judgment, all aspects that, according to the Gigerenzer school, should fog your mind. Results such as this do not mean that the natural frequency findings should be rejected, of course. But it does seem safe to conclude that natural frequency formats have their effects primarily by making set relations transparent, and that they are perhaps the most effective presentational tool for doing this.

The outside view, then, is to think about the two sets, a and $a+b$. The inside view is to focus on the characteristics of the single case, the patient, accident, crime or suspect in question. When you take the inside view, you lose sight of the essential information about sets or classes and the relation between them. Here are some other aspects of thinking where this inside/outside conflict seems to be operating, beginning with one with which you are sure to be familiar through personal experience.

The planning fallacy

Think about a project that you are about to engage in, such as writing a piece of coursework, fixing a car or redecorating a room. How likely do you think it is that

you will complete the task on time? Now think about similar projects that you completed recently. Did you bring them in on time? Research and common experience show us that our predictions tend not to be matched by reality: we are much less likely to meet deadlines than we think we will be before we start a project. We are also likely to underestimate the costs, problems and effort involved. This is the planning fallacy.

Once again, the early running in identifying and researching this problem was made by Kahneman and Tversky (1979a; reprinted 1982c); they even seem to have coined the term 'planning fallacy'. Its essence, as Buehler, Griffin and Ross (2002) explain, is the clash between an overly optimistic prediction about the project in hand and a more realistic history of past performance. Kahneman and Tversky refer to these as the *singular* and *distributional* aspects of planning. These correspond to the inside and outside views. Thus when estimating completion time for a project, we tend to focus on the ways in which this project is unique and forget about the ways in which it is similar to things we have done before.

Of course, there is more to it than just this, because our planning predictions tend to be inaccurate in one direction: overoptimistic. Buehler et al. identify two possible reasons for this. Firstly, planning is by nature about the future, and this orientation may prevent you looking back into the past. Secondly, when you plan a project you plan for its successful completion, not its failure; you will therefore think more about those factors that are likely to promote its success and ignore those that might undermine it – even if you have plenty of experience of them. Buehler, Griffin and Ross (1994) collected verbal reports from students estimating when they would complete an assignment (only 30% met their predicted time). Three-quarters of their collected thoughts concerning these projects were about the future; only 7% referred to their own past experience, another 1% to others' and only 3% referred to possible snags.

Reports of the adverse impact of the planning fallacy are legion, and it can be very costly, indeed lethal. Buehler et al. (2002) quote the case of the Sydney Opera House as 'the champion of all planning disasters' (p. 250): in 1957, it was estimated that it would be completed in 1963 at a cost of $7 million; it opened 10 years late at a cost of $102 million. On an individual level, every year we hear of tourists who think they can get to the top of a mountain and back by tea-time, in their t-shirts and shorts, and are proved wrong in the most drastic way possible: they wind up dead.

How then can we escape the fallacy? Buehler and colleagues put forward three possibilities. The first is to think about sets, that is, to take the outside view; Kahneman and Tversky (1979a) also recommended this remedy. However, this seems hard to do, for a variety of reasons. For instance, Griffin and Buehler (1999) had people make planning predictions about either a single project or a set of ten projects. They contend that the frequentist school of Gigerenzer and his associates would predict that the latter should have reduced the planning fallacy. But it did not: the fallacy was committed to the same extent in both conditions. According to Buehler et al. (2002), this failure was because people were unable to detach themselves from the inside view even when thinking about frequencies. The richness of real-world problems masks their statistical aspects. Taking the outside view is difficult because you have to see the current project as a sample from a population, which, as we saw above, is a sophisticated skill not available to untutored people;

and you have to construct a reference class of similar cases in the past. Sometimes, as Kahneman and Tversky remark, the case at hand is all you have, as in the planning of an opera house.

Secondly, you could think about alternatives to success: how might the project come unstuck? Although this kind of alternative scenario planning is popular in business circles, Buehler and colleagues found, as with the previous remedy, that it failed to overrule the planning fallacy. Even when instructed to generate pessimistic possibilities that were highly plausible, people neglected them in favour of optimistic scenarios. They did, thankfully, find more encouraging support for their third option, which they called the *recall-relevance* manipulation. That is, people were encouraged to focus on their past planning history and think about what it might imply for their prospects of current success. This significantly reduced the gap between prediction and reality.

Finally, it is worth recording aspects of the planning fallacy that Buehler and colleagues point to as a positive advantage for researchers interested in human judgment. It is a clear and simple test-bed for this work because it does not require any reference to logical or mathematical norms, since 'accuracy and bias can be measured by the calendar and the clock' (Buehler et al., 2002, p. 270); it has obvious real-world implications; and it concerns a central psychological question, about how we use memory to guide our thoughts and actions.

Overconfidence

This is another kind of optimistic bias, and has been subject to a lot of attention from researchers: indeed, Griffin and Brenner (2004) call it 'the poster child of judgmental biases' (p. 180). It is related to the calibration of probability judgments, which was discussed earlier. The overconfidence effect is commonly observed with general knowledge questions, such as:

1 Absinthe is: (a) a precious stone; (b) an alcoholic drink.
2 Which city is further north, Paris or New York?
3 Which city has the higher population: (a) Bonn; (b) Heidelberg?
4 What is the capital of New Zealand: (a) Auckland; (b) Wellington?

You can try this for yourself: take each question and, when you have an answer, rate how confident you are that the answer is correct, on a scale of 50–100%. The scale starts at 50% because it is possible to get half these questions right if you know nothing at all about them, just by guessing; 100% means you are certain that you are right.

As long ago as 1980, it was possible to sum up an already large literature on studies of the relation between people's reported confidence and their actual performance. Lichtenstein et al. (1982), in their review of calibration studies, reported that on questions where people said they were 100% confident, they tended to be only about 80% accurate; with 90% confidence, they were 75% accurate, and so on. Their confidence exceeded their performance, just as it does with the planning fallacy.

It did not take long to discover that there is more to this bias than this, and in fact 'overconfidence' may be a misleading term for the overall effect. One early complication was the *hard–easy effect*. Lichtenstein and Fischhoff (1977) gave people three kinds of tests designed to differ in their difficulty: there was an easy test (people got 85% of these questions right), a difficult test (61%) and a test with impossible questions (51% – close to the guessing rate of 50%). There was maximum overconfidence with the impossible test: no matter what the participants thought was their probability of correctness, its actual average rate was always around 50%. The usual overconfidence effect was found with the difficult test: a declining slope from 100% confidence (around 75% accurate) down to 60% (around 55% accurate). However, with the easy set, there was evidence of *underconfidence* at the lower levels: when people were only 50% confident, they were around 60% accurate. They were overconfident when they expressed a high degree of confidence.

Evidence that overconfidence is influenced by more than difficulty was provided by Griffin and Tversky (1992). They distinguished between the *strength* of evidence and its *weight*. In the general knowledge case, strength is your impression about what you know about, say, geography questions such as 2–4 above. Weight is how well that evidence predicts performance. To give another example, you get a very strong tip from someone about the stock market or a horse race, but they know very little about these things (we all know people like that): their advice is strong, but it carries little weight.

Griffin and Tversky gave US students pairs of American states and asked them to choose which one scored more on various attributes: the states' populations, their voting rates in presidential elections and their high-school graduation rates. They predicted that people would be quite accurate and highly confident with the population questions and quite inaccurate and less confident with the voting questions – they were. Most interestingly, they used the graduation question to separate strength and weight: people are likely to have a strong impression that one state is more 'educated' than another, perhaps through availability of well-known universities, but such factors have little predictive validity with respect to school attainment. So there should be low accuracy and high confidence on these questions (i.e. a high degree of overconfidence), which is exactly what Griffin and Tversky found.

The inside–outside dynamic also turns out to be relevant. Several researchers, most notably Gigerenzer and his colleagues, have pointed to the difference in categories between the assessment that you give of your confidence in an answer and the data on your accuracy. The first is a single-case judgment, while the second is a frequency. What would happen if the confidence judgment were also asked for in frequency terms? Gigerenzer, Hoffrage and Kleinbölting (1991) addressed this question. Instead of asking people to rate their confidence in single questions, they asked them to assess how many of the whole set of test questions they thought they had got right. If people were truly overconfident, this should make no difference. However, Gigerenzer et al. found that the overconfidence effect disappeared with this frequency task. In fact, it went into reverse: in one experiment, there was a 14% overconfidence effect when it was tested for in the normal way, with single questions; but with the frequency question, there was a difference between confidence ratings and accuracy of more than –2%, indicating a slight degree of underconfidence.

Gigerenzer has a particular explanation for this effect, among others, which we shall return to in Chapter 9 since it relates to decision making as well as judgment. For now, we can note that we have again what appears to be an inside–outside difference to do with the reference classes that the two types of task, single-question and whole-test rating, invite you to invoke. In the case of the single question, you might ask yourself what you know about questions like that particular one: geography, say, or weird drinks. With the whole test question, you ask yourself about how well you tend to do on general knowledge quizzes. As with the planning fallacy, we tend to be better calibrated when it comes to our past histories than we are when we need to apply them to the case in hand.

The conjunction fallacy

This is also a very well researched aspect of thought, and one that brings the inside–outside distinction into sharp focus. It involves possibly the most famous fictional character in cognitive psychology. She is called Linda, and this is the Linda problem:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-war demonstrations.
>
> Which descriptions are most likely to be true of Linda? Rank them in order of probability.
>
> *a* Linda is a primary school teacher.
> *b* Linda works in a bookshop and takes yoga classes.
> *c* Linda is active in the feminist movement.
> *d* Linda is a psychiatric social worker.
> *e* Linda is a member of the League of Women Voters.
> *f* Linda is a bank clerk.
> *g* Linda sells insurance.
> *h* Linda is a bank clerk and is active in the feminist movement.

As always with these problems, it will do you good to have a go at it before reading on.

The original version of the problem was reported by Tversky and Kahneman (1982b, 1983); I have altered the wording slightly. Linda appeared alongside another problem involving Bill, but he lacked Linda's charisma and has been largely forgotten. Now look at your rankings, and especially at the numbers you wrote against sentences *c*, *f* and *h*. These are the important ones; the others are merely fillers.

If you are like most people given this task, you ranked *c* above *h* and *h* above *f*. If you did the second of these, as 89% of participants did in an experiment reported by Tversky and Kahneman (1983), you have made a logical error. Why? Because it is impossible for a conjunction, $x + y$, to be more probable than one of its

components. You cannot be more likely to be a female student than just a student, or a feminist bank clerk than just a bank clerk. Note that people tend not to judge that Linda is more likely to be a feminist bank clerk than just a feminist; they rank *c* above *h*. This gives a clue as to what is going on here. Tversky and Kahneman interpreted their findings in terms of the representativeness heuristic, which was discussed earlier: one of the aspects of this heuristic was that people judge an item, such as sentence *f*, as likely to the extent that it is similar in essential properties to its parent population. Linda sounds a bit of a radical, doesn't she? The parent population then is something like 'radical women'. So people think that it is more likely that she is a radical bank clerk than just a bank clerk: sentence *f* might imply that she is *not* a feminist (Politzer & Noveck, 1991).

Thinking about Linda in this way is taking the inside view and focussing on her individual attributes. What happens when people are asked to take the outside view, and focus instead on the relation between classes? This was in fact done in the Tversky and Kahneman (1983) paper, and a replication, using a slightly different method, was reported by Fiedler (1988). Fiedler used a ranking task, as in the example above, whereas Tversky and Kahneman had people give probability ratings to each item. Hertwig and Gigerenzer (1999) brought the natural frequency theory to bear on the problem. In essence, what all these studies did was to ask participants to think about 100 people just like Linda, and then assess statements *f* and *h*. In all cases, the conjunction fallacy was committed far less often, sometimes disappearing altogether.

Sloman and Over (2003) present evidence that, once again, it is not frequency presentation itself that brings about this insight, but its ability to make set relations transparent: in this case, people need to see that $x + y$ is a subset of x. They found that this facilitation could be suppressed if the set relation was obscured by having the critical statements (*f* and *h* above) separated by seven filler items rather than just one. This happened with both rating and ranking tasks, and they also found, as previous researchers had, that the ranking task produced much higher rates of the fallacy; this point is discussed in detail by Hertwig and Chase (1998), as well as by Hertwig and Gigerenzer and Sloman and Over. A highly detailed review of the conjunction fallacy, covering more theoretical interpretations of it than I have space to include here, is given by Fisk (2004).

We have seen in this chapter how crucial probability judgment is in real-life thinking, and as stated at the outset it is also crucial in explaining human reasoning: it is central to the 'new paradigm' view of reasoning that has recently emerged from the schools of Oaksford and Chater and Evans and Over (Over, 2009). It is also at the core of two other forms of thinking that we shall deal with: inductive thinking and decision making. So let us first delve into the history of the study of reasoning, and from there proceed to the new paradigm. Then we shall head off into these other areas.

Summary

1 Probability is at the heart of the explanation of many areas of thinking and reasoning. It is a research topic in itself, and has always been a cornerstone

of theories of decision making, but has recently come to the fore in the psychology of reasoning as well.

2 Probability can be formally defined in four ways: as logical possibility, frequency, propensity or subjective degree of belief.

3 People's judgments of probability may systematically depart from mathematical norms: they have an insecure understanding of randomness, and tend to overweight low probabilities, among other biases.

4 Sampling is a major problem in naïve probability judgment: people are unaware of sampling biases and how they can bias judgment.

5 The most influential school of thought in the psychology of probability judgment is the *Heuristics and Biases* research programme of Kahneman and Tversky. They explain sampling biases through the availability heuristic.

6 Belief updating is the revision of beliefs in response to evidence. Its normative theory is derived from Bayes' rule, which weights prior belief by the conditional probability of the evidence, given that belief and its alternatives, to derive a posterior probability.

7 Kahneman and Tversky explain deviations in human performance in belief updating through heuristics such as representativeness.

8 Biases such as base rate neglect have been explained by Gigerenzer using the construct of natural sampling, that gives rise to representations of frequencies that obviate the need to consider base rates. This approach has been successfully applied to improve judgment in medical and legal contexts.

9 Recent research implies that the natural sampling effect may come about through its making set relations transparent. It facilitates an 'outside' view of probabilities (the set of cases) while bias often results from taking an 'inside' view (of the case at hand).