2 Quantifying uncertainty

DAVID SPIEGELHALTER

Putting numbers on risks

Risk is a strange concept. Different disciplines have tried to define it precisely, but perhaps it is better to be informal and follow more popular usage. I shall take it as *anything to do with situations where 'bad' (or 'good') things may, or may not, happen.* The crucial elements are that there is uncertainty, and that the outcomes may be nice or nasty.

A wealth of recent psychological research has shown that we mainly use 'gut feelings' to deal with such situations, rather than carefully weighing up the consequences and assessing numerical probabilities, as more formal approaches would have us do. Our feelings are influenced by culture, our experiences and those of people close to us, media coverage, emotional feelings of dread, or hope, and so on, but we manage to get by most of the time, and it is noticeable how recently, in historical terms, the theory combining probability and 'rational' decision-making was developed. Even when evidence is available about the 'size' of a risk, in sufficiently stressful situations it may be ignored. Cass Sunstein, a senior adviser to Barack Obama, claims that people display 'probability neglect' when confronted with vivid images of terrorism, so that 'when their emotions are intensely engaged, people's attention is focused on the bad outcome itself, and they are inattentive to the fact that it is unlikely to occur'. So the 'true' risks are ignored; it's been shown that people are, rather illogically, willing to pay more for insurance against terrorism than insurance against all risks (which implicitly include terrorism), just because the use of the word conjures up dread.

But gut feelings might be unreliable in some circumstances, for example when people are trying to manipulate you to take some action, or when

https://doi.org/10.1017/CBO9780511735950.002 Published online by Cambridge University Press

the reasoning is complex and a lot depends on the decision. Then a more analytic, and perhaps rather unnatural, approach can be useful, whether you are an individual trying to make a personal decision, or you represent an organisation or government deciding on a policy in the face of uncertainty.

This more formal approach relies on putting numbers on probabilities of events, and raises the inevitable question: *can we quantify our uncertainty?* In this chapter we will just look at this question, ignoring our knowledge and feelings about the consequences of actions.

Putting probabilities on events

In some circumstances we can use pure logic to come up with reasonable probabilities, because of the assumed symmetries in the situation which allow equally likely outcomes to be specified. These are the classical areas of probability, with balanced coins, shuffled cards, and so on. For example, in the UK National Lottery six balls are drawn without replacement from a drum containing forty-nine numbered balls. If the numbers match the six numbers on your lottery ticket then you win, or share, the jackpot – fewer matches win less, with the lowest prize being ten pounds for three matching numbers.

If we assume that the lottery-drawing mechanism is completely fair and unbiased, so that each number is equally likely to be drawn, then we can immediately calculate the probabilities of specific events, such as a 1 in 13,983,815 chance of winning a jackpot, and a 1 in 56 chance of matching three numbers. Note the use of the word 'chance', deliberately carrying the connotation of an 'objective' number that can be calculated using the theory of probability.

If these probabilities are *assumed* to be known, because of the physical properties of the system, then we can learn nothing from history – even if the same lottery numbers came up every week we would have to put it down to luck. But even the slightest suspicion of irregularities changes everything, and suddenly the reassuring calculations evaporate if, for example, you suspect that some of the balls have been left out of the bag. The vital conclusion is that these 'classical' probabilities – chances that are states of the world – are grounded on

subjective assumptions about the generating mechanism, and hence are deeply contingent.

An alternative basis for quantifying uncertainty is by using historical data. If the future follows the same pattern as the past, then frequencies of events in history should reflect reasonable probabilities for events in the future. In fact the 'frequentist' view defines probability as the limiting frequency in a (fictitious) infinite replication. For example, sports betting companies use past data on football matches to propose reasonable odds for the results of future games. We have tried this, using fairly simple models involving estimates of the 'home advantage', 'attack strength' and 'defence weakness' of teams, that can be combined to give us expected numbers of goals in each match and hence, assuming a Poisson distribution around the expected values, a probability for any particular final score. Our own models have met with mixed success, but the (confidential) models used by professionals presumably work well enough to make money.

The assessed probabilities are therefore based on assumptions about the continuity of past with future, together with an assumed mathematical model for how various fictitious parameters such as 'attack strength' interact to give rise to appropriate odds. The lesson from this kind of exercise is that such assumptions are *known* to be false, or at least not precisely *true*, and yet the resulting probabilities may be good enough for the purpose. Again, the final numbers are contingent upon unprovable assumptions.

Finally, the situation may have neither reassuring symmetries nor useful historical precedents. For example, consider the situation in early 2008. The probability of Barack Obama becoming the next President of the United States could hardly be based on the empirical historical record of forty-three out of forty-three US Presidents being white. Philosophically, we might believe there is still some objective 'propensity' in the situation for Obama to be the next President, but this does not seem practically useful. Instead we are left to make a judgement using existing information, expressed as the betting odds that we are willing to place or to lay bets. These lead to the results shown in Figure 2.1, which are derived from a major online betting exchange. These 'probabilities' are not based on any 'objective' state of the world, nor historical data, and change

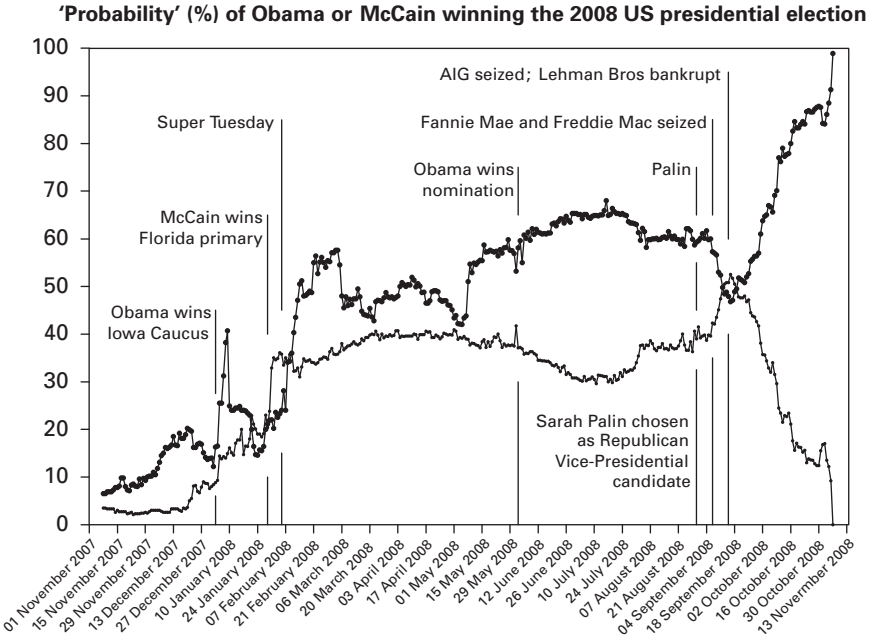

FIGURE 2.1 'Probability' of Barack Obama (dark line) or John McCain (light line) winning the 2008 US Presidential election, as reflected in the odds both taken and offered on a betting exchange, each day in the year up to the 2008 election.

constantly in receipt of further information. I would still argue that these are reasonable probabilities, as they reflect reasonable numerical uncertainty concerning the outcome, given the current state of knowledge.

These three circumstances – classical symmetry, historical data and subjective judgement – all lead to precisely the same conclusions. Probabilities are constructed on the basis of existing knowledge, and are therefore contingent. This rather dramatic conclusion, although open to dispute by some statisticians and philosophers of probability, has a respectable pedigree among the community of Bayesian statisticians. Indeed a guiding quote throughout my career comes from Bruno de Finetti:

> Probability does not exist.

https://doi.org/10.1017/CBO9780511735950.002 Published online by Cambridge University Press

I take this to mean that probabilities are not states of the world (except possibly at the sub-atomic level, about which we are not concerned here), but depend on the relationship between the 'object' of the probability assessment, and the 'subject' who is doing the assessing. This means that, strictly speaking, we should not use the phrase '*the* probability *of* X', but '*my* probability *for* X', where 'my' refers to whoever is taking responsibility for the probability. This makes clear that probability expresses a relationship, not a property or an objective fact about X. Sadly, this phrasing is unlikely to become standard practice.

The second guiding quote for my career comes from a great industrial statistician, George Box:

> All models are wrong, but some are useful.

Again, this emphasises that the mathematical structures that we construct in order to arrive at numerical probabilities are not states of the world, but are based on unprovable assumptions. We shall look briefly at the deeper uncertainties concerned with model-building in a later section.

Representing probabilities

There is a wide range of alternatives for representing probabilities when communicating with different audiences. Here we discuss a limited list of options, and briefly summarise some of the psychological research related to the perception of the magnitudes of probabilities associated with the different representations.

By putting my personal information through a computer program, for example, my general practitioner can tell me that I have around a '10 per cent chance' of a heart attack or stroke in the next ten years. How might such a quantity be communicated to me? We first consider the use of text, either using words or numbers:

- **Natural language**: For example, 'you *might* have a heart attack or stroke', or 'it *is possible* you . . . '. Such language is widely used in weather forecasting. The interpretation of such terms is highly dependent on the subject: if numerical information is to be communicated, a fixed 'translation' might be agreed, such as in recent Intergovernmental Panel

for Climate Change (IPCC) reports in which 'very likely' is taken to mean more than 90 per cent confidence in being correct.

- **Numerical probabilities defined between 0 and 1**: For example, 'Your probability of having a heart attack or stroke is 0.1.' This format is never used except in technical discussions.
- **Numerical 'chances' expressed as percentages**: 'You have a 10 per cent chance . . . '. This is widely used in popular discourse, but has connotations of random devices such as dice, which can appear inappropriate when discussing serious personal issues.
- **Numerical 'odds'**: 'You have a 1 in 10 chance of . . . '. This is a more popular expression, although it is still in terms of chances, but means that smaller probabilities are associated with larger numbers. Recent evidence suggests that around 25 per cent of the adult population cannot say which is the largest risk out of the options '1 in 10', '1 in 1000', '1 in 100'.
- **Frequencies in populations:** For example, 'Out of 100 people like you, 10 will have a heart attack . . . '. This is becoming a common text representation in leaflets and computer programs designed to explain risks to medical patients. However, it requires one to see oneself as part of a group of similar people – a 'reference class' – and this could conflict with a self-image of uniqueness and lead to a denial of the relevance of the statement.
- **Frequencies out of 'possible futures':** 'Out of 100 ways things might turn out for you over the next 10 years, you would be expected to have a heart attack or stroke in 10 of them.' This is a novel representation intended to encourage the immediacy and ownership of the risk. Philosophically it is very shaky: it is an uneasy mixture between a probability, constructed on available knowledge, and a frequency interpretation, as a proportion of a population of possible futures.

There is also a range of graphical options, including pie charts, circles representing the size of the risk, bar charts, icon-plots showing many small 'people', 'Smilies' showing multiple iconic faces experiencing different outcomes, multiple photos, and word-clouds, in which the size of the font is proportional to the probability of the event.

None of these presents a universal solution. Challenges that arise include:

1. Comparing rare and more common events, leading to the frequent use of graphics on a logarithmic scale, or providing a 'magnifying glass' for zooming in on rarer events.
2. Graphical representation of multiple outcomes for the same individual.

https://doi.org/10.1017/CBO9780511735950.002 Published online by Cambridge University Press

3. Comparisons between alternative options when each brings a mixture of potential harms and benefits.
4. Uncertainty, in the sense that we may be more confident about some probabilities than others. While in principle this is of limited relevance, and may only add additional complexity, it could be an option for more sophisticated users.

All these formats deal with probabilities of single events, rather than uncertainty about continuous quantities such as future income, where a wide range of additional graphical tools would be necessary. Uncertainty about the time until an event, such as death, requires a representation for the distribution of possible survival times, for which there is a further range of options which are not explored here.

When it comes to evaluating different formats, it is important to be clear about the purpose of the representation. Broadly, we can divide the aims into:

- Gaining immediate attention and interest.
- Communicating information to be retained.
- Influencing continuing behaviour.

There are clear similarities between these objectives and those of commodity and service marketing. It would be intellectually satisfying to find that the three objectives follow a nice causal pathway: gaining interest leads to knowledge retention which influences behaviour. However, the research literature suggests that the relationship between these objectives is complex, if it exists at all. We must therefore be clear about what we are trying to achieve. For shared-care decisions in health, for example, we may want to provide information so that everyone feels they have made an informed choice, but without necessarily directly trying to influence behaviour in one direction or another. Research suggests that many people strongly welcome information provision, but then seek a fairly paternalistic form of advice ('I really appreciate you telling me all this, doctor, but what do you think I should do?')

There are possibilities for combining many of these representations within a single interface using interactive animations. Given that there is no single 'best' representation, this seems an appropriate policy so that users can essentially choose, or be guided towards, the format that they

find most natural and comprehensible. And of course we must remember that people are likely to be much more influenced by their trust in the information source, their personal experiences and their feelings about the possible outcomes, than they are by the particular choice of format. Nevertheless, it seems a reasonable duty to try and do our best to make sure that the available evidence is permitted to play a role in personal decisions.

Communicating small lethal risks

There are particular problems in comparing and communicating small lethal risks, and yet this is what many of us are faced with in our daily lives. Ideally we need a 'friendly' unit of deadly risk. A suggestion made in the 1970s by Ronald Howard is the use of the *micromort*, or a one-in-a-million chance of death. This is attractive as it generally means that we can translate small risks into whole numbers that can be immediately compared. For example, the risk of death from a general anaesthetic (not the accompanying operation), is quoted as 1 in 100,000, meaning that in every 100,000 operations we would expect one death. This corresponds to ten micromorts per operation.

We can also consider the 18,000 people out of 54 million in England and Wales who died from non-natural causes in 2008, such as accidents, murders, suicides, and so on. This corresponds to an average of $18{,}000 / (54 \times 365) \approx 1$ micromort per day, so we can think of a micromort as the average 'ration' of lethal risk that people spend each day, and which we do not unduly worry about. A one-in-a-million chance of death can also be thought of as the consequences of a bizarre game in which twenty coins are thrown in the air, and if they all come down heads the thrower has to commit suicide. It is interesting to explore, in a fictitious context, the amount people would accept as payment to take part in the game.

A measure such as a micromort needs a unit of exposure to accompany it, and different sources of risk naturally give rise to different levels of exposure. Here we briefly consider transport, medical events, and leisure activities. Of course, we can only quote 'average' risks over a population, which neither represent 'your' risks nor necessarily those of a random person drawn from the population. Nevertheless they provide useful 'ballpark' figures from which reasonable odds for specific situations might

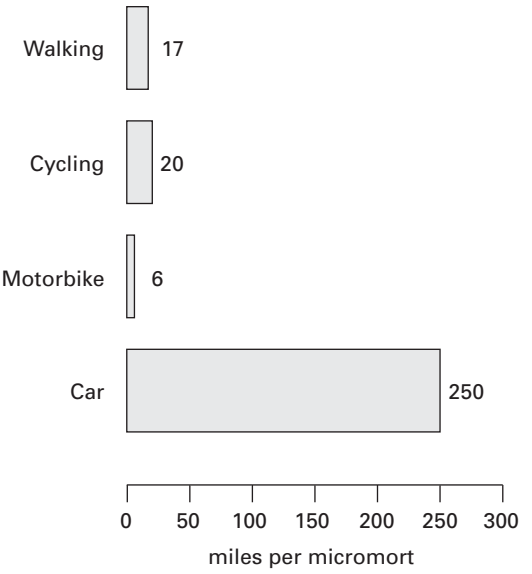

FIGURE 2.2 Distance travelled per micromort (one-in-a-million chance of death) for different forms of transport in the UK, based on assumption of constant risk within transport type and over time.

be assessed. As we have already emphasised, we do not consider that numerical risks exist as fully estimable properties of the world.

Transport

Options for comparing forms of transport include miles per micromort, micromort per 100 miles, micromorts per hour, and so on. We compare the first two options below. Although the general advice is that larger numbers should correspond to larger risks, 'miles per micromort' seems attractive, especially when used to provide a 'calibration' against other risks.

We have not included trains and planes as they would require a change in axes, and the rarity of fatalities (even though they are given great coverage) makes assessment of 'average' risk of limited value.

Medical events

Since these are specific, discrete occurrences, the natural measure is risk per event, for example giving birth, having a Caesarean section or having a general anaesthetic. The exception is spending time in hospital, which

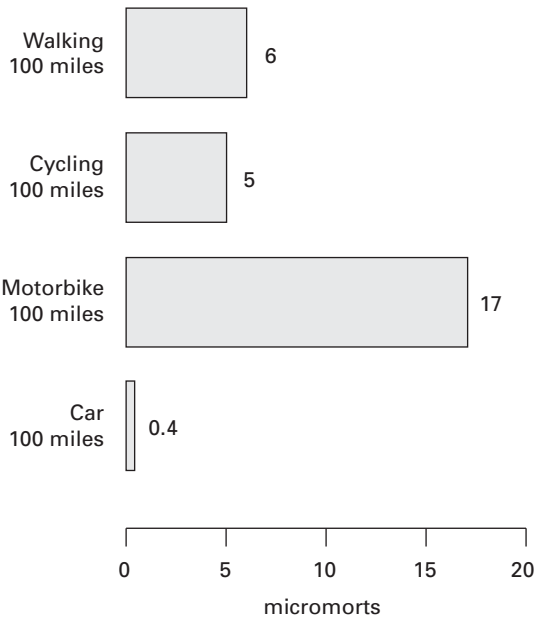

FIGURE 2.3 Micromorts per distance travelled for different forms of transport in the UK, based on assumption of constant risk within transport type and over time.

is expressed as micromorts per night spent in hospital, considering only deaths that were not due to natural causes.

Examination of Figures 2.2 and 2.4 is informative. For example, having a general anaesthetic carries the same risk of death, on average, as travelling 60 miles on a motorbike. The high value for a night in hospital is derived from the National Patient Safety Agency reports of adverse events resulting in death. If anything, this is an underestimate.

Leisure activities

We assume that the risk comes from a specific 'accident' in what is an otherwise safe activity with no chronic ill effects. It is therefore natural to express exposure as the specific activity. Since the activities take different lengths of time it would be possible to express them as micromorts per hour, but this does not seem to reflect the choices that people make.

All these examples concern sudden deaths, but many 'risky' behaviours have a delayed impact, such as smoking or an unhealthy diet. Comparing

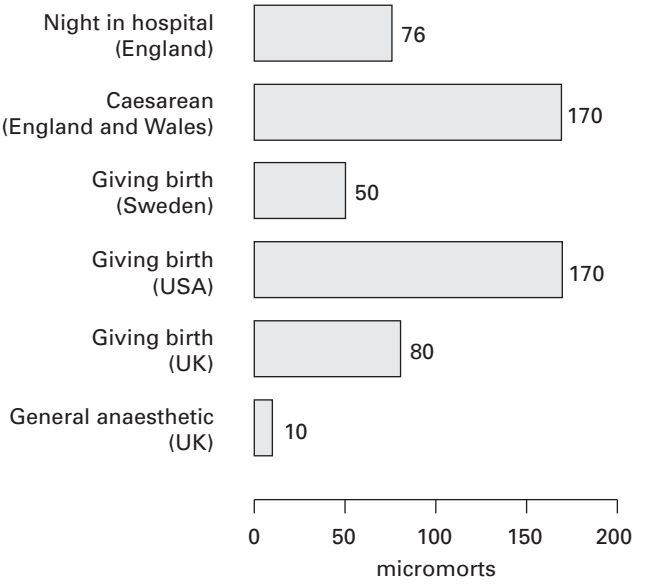

FIGURE 2.4 Micromorts per medical event in 2008, based on assumption of constant risk within event type and over time.

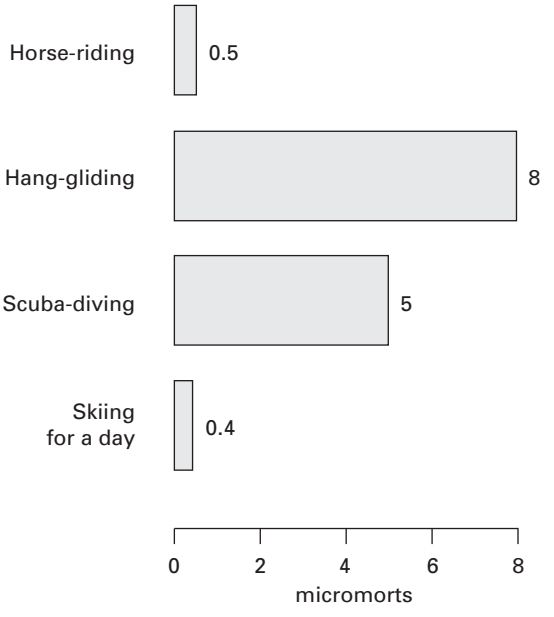

FIGURE 2.5 Approximate micromorts per activity, based on assumption of constant risk within activity and over time.

https://doi.org/10.1017/CBO9780511735950.002 Published online by Cambridge University Press

'acute' and 'chronic' risks is tricky, but there have been a number of suggestions for 'riskometers' which attempt to put both immediate and delayed risks on a common scale. Options include:

1. Listing causes of death, which allows a comparison of how many people, for example, die from accidents compared to heart disease, but does not directly allow the comparison of alternative daily activities.
2. Transform lifetime risks, say of dying from cancer due to smoking, into risk per day by assuming that there are, say, around 30,000 days in a lifetime. But no allowance is made for delayed effects.
3. Discount future risks by a specified factor, and assess the proportional loss on discounted life-expectancy due to different activities. This can then be converted to a logarithmic scale.

None of these options seems entirely satisfactory, as they inevitably mean placing, for example, cigarette smoking and motorcycle riding on a common risk scale, and yet these two behaviours have very different consequences.

Epistemic uncertainty

We have seen how the theory of probability is used as a tool in analysing the essential unpredictability of existence, also known as *aleatory* uncertainty. For example, before I flip a fair unbiased coin, people are generally willing to say there is a 50 per cent chance of a head. However, if I flip the coin and then cover up the result, and ask what is the probability of a head, after some grumbling an audience may be willing to admit the odds are still 50:50. Their misgiving is understandable – they need to cross an important line in being willing to use numerical probabilities to express their *epistemic* uncertainty, that is, their ignorance about what the coin actually shows. What then if I look at the coin and ask them the for probability of a head? After an even longer pause they may grudgingly admit it is still 50:50. Now they have been dragged into the full recognition that epistemic uncertainty is not a property of the object, in this case the coin; it is a property of their relationship with the object, and we all may have different epistemic uncertainties depending on the knowledge to which we are privy. This is the essence of the Bayesian approach to statistics. It allows us to use probability theory to express epistemic uncertainty.

Table 2.1 *Scoring procedure when expressing your confidence as a probability of being correct*

Your 'probability' that your answer is correct (out of 10)	***5***	***6***	***7***	***8***	***9***	***10***
Score if you are ***right***	0	9	16	21	24	25
Score if you are ***wrong***	0	−11	−24	−39	−56	−75

Of course a problem arises when we are deluded about the knowledge we have and claim certainty, or at least high confidence, in facts that are not actually the case. Fortunately there is an under-appreciated branch of statistics concerned with assessing the quality of people's probability judgements using what are known as scoring rules. These are designed to penalise people for providing poor probabilities for events whose truth or falsity is later determined.

We can illustrate the issues with some simple questions given below. In each case either (A) or (B) is the correct answer, and the challenge is to decide which answer you feel is most likely to be correct, and quantify your probability that your answer is correct. So if you are certain (A) is correct then you should give it 10/10, but if you are only around 70 per cent sure then it gets 7/10. If you have no idea, then give 5/10 to either choice.

1. Which is higher: (A) the Eiffel tower, or (B) Canary Wharf?
2. Who is older: (A) George Osborne, or (B) Nick Clegg?
3. In the International Movie DataBase rankings (29/12/2009), which film comes higher: (A) *The Matrix*, or (B) *Forrest Gump*?
4. Which is bigger: (A) Croatia, or (B) Czech Republic?
5. Which is bigger: (A) Venus, or (B) Earth?
6. Who died first: (A) Beethoven, or (B) Napoleon?

Table 2.1 shows how you are scored when the true answer is revealed. If you are absolutely correct then you score twenty-five, but if completely wrong then you lose seventy-five. If your probability was five for either answer, then you stay where you were. It is clear that there is a steep penalty for being confident and wrong. This is not arbitrary punishment, but a consequence of designing a scoring rule that encourages honesty, so that if you are 70 per cent sure of, say (A), then your expected score

is maximised if you give a probability of 7/10 for (A), rather than exaggerating and giving a probability of 10/10 for (A). Such a scoring rule is called 'proper'.

By subtracting twenty-five from each of the scores it becomes clear that the penalty is dependent on the square of the probability given to the wrong answer. This quadratic, or Brier, scoring rule was developed to train weather forecasters to give reasonable probability of future weather events such as rain. A simple linear scoring rule, such as scoring somebody by the probability given to the correct answer, is inappropriate as it would encourage people to exaggerate their confidence in being right.

This process shows that epistemic uncertainties can be quantified as probabilities, which are necessarily subjective and expressed by an individual on the basis of available knowledge. They should not be thought of as embodying some 'true belief', but are *constructed* by whatever elicitation process is being used. But for these judgements to be useful, people's probabilities need to have some reasonable properties. First, they should be *calibrated*, in the sense that if someone gives a probability of 7/10 to a series of events, then, around 70 per cent of those events should actually occur. Second, the probabilities should *discriminate*, in that events that occur should be given higher probabilities than those that do not. It can be shown that a proper scoring rule rewards both calibration and discrimination.

So far we have considered events that are well defined and whose truth can be established. In real situations things are generally not so simple, as we shall explore in the next section.

Deeper uncertainties

In 1921 Frank Knight published his book *Risk, Uncertainty and Profit*, in which he distinguished between 'risk' and 'uncertainty'. 'Risks' were objective quantities that could either be obtained by reasoning (for example, symmetric situations involving dice, cards, etc.), or estimated from historical data. Conversely, 'uncertainty' was subjective and judgemental, and not susceptible to objective measurement. Since that time the use of subjective probabilities has become developed and so, as our discussion in earlier sections shows, it may be considered reasonable to

use numbers to express our subjective beliefs. However, there will still be many circumstances in which we feel that our ignorance is so great, or the future possibilities so ill-defined, that we are unwilling to express numerical judgements with any confidence.

We may also have the courage and insight to acknowledge that there may be important things we have not even thought of like the Rumsfeldian 'unknown unknowns'. A famous plea for such humility came from Oliver Cromwell. In 1650 he was trying to avoid a battle with the Church of Scotland, which was then supporting the return of Charles I's son. He wrote: 'Is it therefore infallibly agreeable to the Word of God, all that *you* say? I beseech you, in the bowels of Christ, think it possible you may be mistaken.'

If we are willing to entertain the possibility that we may be mistaken, then it may mean we have crossed the border of quantifiable uncertainty, and opened up the possibility of non-numerical expressions of our doubts and ignorance after we have constructed a model from which we want to derive risk assessments. These misgivings may take many forms. For example, we might conduct analyses under alternative sets of assumptions, and examine the robustness of our conclusions. We may admit to aspects of the world that we know have not been adequately included, and informally express our judgement as to their importance. We may express judgements as to the strength and quality of the evidence underlying our model and so express limits to our confidence in some conclusions. We may add on a 'fudge factor' to allow for all the things we may not have thought of. Finally, we may, of course, choose to deny non-modelled uncertainty, or unwittingly overlook errors in our model. One can see examples of these strategies being played out in the deliberations about climate change.

Conclusions

We have shown how our uncertainties about events can be quantified using probability theory, whether or not there is a firm logical or historical basis for these assessments. By taking a Bayesian perspective, we can extend the use of probability to cover our epistemic uncertainties about well-defined quantities. We may even, in some circumstances,

quantify our uncertainty about the appropriate model to use. But when we start acknowledging our inability to represent the full complexity of the real world using mathematical models, we are faced with leaving the safe land of quantifiable uncertainty and entering the (possibly hostile) environment of disputed science, ill-understood possibilities and deep uncertainty.

This is a world in which many statisticians and mathematical scientists feel very uncomfortable, and for which they receive no training. A good start to their education might involve acknowledging that their models are inadequate constructs derived from currently accepted knowledge, and that numerical probabilities are not a property of the world but an expression of their subjective understanding of the world. These may be considered fairly radical ideas.

On the other hand, those tasked with taking action on the basis of risk assessments derived from a formal model also need to accept their provisional and contingent nature, and the associated deep uncertainties. This they may be reluctant to do, in their desire for concrete guides on which they can base decisions.

My own feeling is that, when decision-makers are dealing with 'expert' risk assessments based on models, there should be quantification of uncertainty to the maximum possible extent. But the potential limitations of these numerical assessments should be acknowledged. A language is also required for communicating, with due humility and without fear of casual rejection, the deeper uncertainties.

Answers to quiz

A, B, A, B, B, B

Acknowledgements

I would like to thank Mike Pearson and Ian Short with help in preparing animations and illustrations for the lecture, Ted Harding for constructing Figure 2.1, and Frank Duckworth for discussions on micromorts and riskometers.

References

Box, G. (1979) *Robustness in Statistics: Proceedings of a Workshop*. New York: Academic Press.

De Finetti, B. (1974) *Theory of Probability: A Critical Introductory Treatment*. London, New York: Wiley.

Galesic, M. and Garcia-Retamero, R. (2010) 'Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples', *Archives of Internal Medicine* 170 (5): 462–8.

Gigerenzer, G. (2010) *Rationality for Mortals: How People Cope with Uncertainty*. New York: Oxford University Press.

Knight, F. (1921) *Risk, Uncertainty and Profit*. Available at: www.econlib.org/library/Knight/knRUP.html.

Lindley, D. (2006) *Understanding Uncertainty*. Hoboken, NJ: Wiley.

Lipkus, I. M. (2007) 'Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations', *Medical Decision Making* 27 (5): 696–713.

Micromort – Wikipedia, the free encyclopedia. Available at: en.wikipedia.org/wiki/Micromort (accessed 23 August 2010).

Morgan, G., Dowlatabadi, H., Henrion, M. et al. 'Best practice approaches for characterizing, communicating, and incorporating scientific uncertainty in decision making'. Final Report, CCSP Synthesis and Assessment Product 5–2. Available at: www.climatescience.gov/Library/sap/sap5–2/final-report/default.htm (accessed 10 July 2010).

Sunstein, C. R. (2003) 'Terrorism and probability neglect', *Journal of Risk and Uncertainty* 26: 121–36.