ELICITING A PRIOR DISTRIBUTION

John Quigley, Tim Bedford and Lesley Walls

Department of Management Science

University of Strathclyde

Glasgow

United Kingdom

ABSTRACT

An overview of key issues associated with the elicitation of a prior probability distribution is provided. The nature of subjective probability is discussed and the main causes of biased assessments identified. Evidence is presented suggesting that people can improve their probability assessments under the correct conditions. The ways in which the elicitation of a prior distribution can be embedded within a process to minimise the impact of bias is examined. More advanced issues, namely validation of a prior distribution and reducing the cognitive burden on experts, are described.

1 INTRODUCTION

Expert engineering judgement plays a vital role in assessing and modelling the quality and reliability of systems (Keeney and von Winterfeldt 1991; Bedford et al to appear), although presenting subjective judgements as formal probability distributions is not without controversy (Evans 1989). Like any data collection process, a sound methodology for elicitation is a pre-requisite to acquiring reliable subjective expert judgment which will be summarised in the form of a prior distribution.

A prior distribution is a subjective probability distribution describing some unknown quantity of interest that has been elicited from a relevant person(s) prior to observing data. Through the use of Bayes's Theorem, the prior distribution is updated in the light of data to

produce a posterior distribution. Updating requires a conditional probability distribution, known as the likelihood function, which describes the probability of observing the data conditional on the state of the unknown quantity of interest.

A subjective probability distribution is one that describes a person's belief in the value of an unknown quantity. Winkler (1967) believes that there is no single correct subjective probability distribution for an unknown quantity because each person forms his/her beliefs from personal experience. Cooke (1991) points out that there is no way of measuring the correctness of a subjective distribution. In contrast, O'Hagan (1988) distinguishes between true and stated probabilities. The former would result if an expert were capable of perfectly accurate assessments, while the latter would result from attempts to specify the underlying true probabilities. The authors agree that a good prior distribution will conform to the laws of probability, although the differing perspective has implications for validation.

The Bayes coherency principal states that uncertainties are described by probabilities and subjective probabilities should be such as to ensure self-consistent betting behaviour (de Finetti 1974). This means that a person assesses subjective probabilities in such a way that they would avoid gambles involving certain losses (a so-called dutch book). In principle subjective probabilities can be measured through preference behaviour, when a subject declares indifference between the event under discussion and a lottery with a known probability (assuming of course that the "prizes" are comparable).

Even if a prior distribution conforms to the laws of probability, some elicited distributions provide more useful inference than others. Research in experimental psychology has demonstrated that accurate subjective probabilities are unobtainable by simply asking someone to provide a probability number; instead a structured elicitation process is required (Kahneman et al 1982, Merkhofer 1987). These processes consist of an 'expert' (or set of experts) and an 'analyst'. Following the definition of Ferrell (1994), an expert is "a person with

substantive knowledge about the events whose uncertainty is to be assessed”, while an analyst should be a suitably qualified individual who will collect the data from the expert in order to formulate the prior distribution for use in modelling. The aim of an elicitation process is to minimise the impact of biases inherent in surfacing and capturing subjective expert judgement.

After examining aspects of modelling relevant to the elicitation of a prior distribution in section 2, section 3 describes the typical biases encountered during elicitation. Section 4 further considers characteristics of good prior distributions, while section 5 highlights issues associated with improving the ability of assessors to provide useful inference. Section 6 outlines a generic process for supporting the elicitation of a prior distribution and references specific processes used in reliability. Finally, section 7 reflects upon advanced theoretical and practical issues concerning the elicitation of prior distributions.

2 MODEL PARAMETERISATION AND UNCERTAINTIES

It is useful to distinguish between different types of uncertainty in modelling: aleatory uncertainty which corresponds to inherent randomness in the process generating observations; and epistemic uncertainty which corresponds to state of knowledge of an expert. Within a Bayesian statistical model the aleatory uncertainty is described by the likelihood function, while the prior distribution describes the epistemic uncertainty. Typically, as more relevant data are acquired, epistemic uncertainty is reduced, although not the aleatory uncertainty. As an example, consider the uncertainty associated with determining the length of time a component will operate without failing. The uncertainty associated with the mean time to failure of a component from a particular class is epistemic and will reduce through observing the times to failure for such components. However, even if the mean time to failure of components within this class is known, then the aleatory uncertainty would still remain with respect to the actual time to failure for any one component.

Many of the favoured model parameters, such as a failure rate and mean time to failure, are not actually observable quantities, but are simply parameters of a model used to make predictions about the future. It has been strongly argued on foundational grounds, for example, the discussion in Chapter 2 of Bedford and Cooke (2001) that an expert should only be asked for probability assessments on observable quantities. This raises an interesting issue concerning the degree to which the model construct relates to an expert's perception and hence the coverage of the elicitation. For example, Bedford et al (to appear) propose that elicitation should include qualitative structuring of the model as well as the quantitative assessment of the prior distribution for the quantity of interest to facilitate appropriate parameterisation.

3 HEURISTICS AND BIASES

Research in psychology is relevant to the elicitation of a prior distribution. The identification of the major sources of bias within expert assessments motivates the need for supportive elicitation processes and validation of data.

Sunstein (2005) discusses two alternative cognitive systems to describe how a person makes decisions: System 1 uses intuition, and results in quick decisions which can be subject to error; System 2 is more deliberative, calculative and less likely to produce errors. System 1 is based on heuristics, whereby a person develops short cut approaches to making decisions. While heuristics can provide insight into a problem, careful deliberation should result in an improved assessment. This implies that the elicitation of a prior distribution requires an analyst to engage the expert in System 2 thinking. Hammond (see O'Hagan et al 2006) describes a person's thinking as a continuum between the aforementioned dichotomy.

Tversky and Kahneman (1974) were the first authors to make substantial contributions to the identification of heuristics used within probability assessments and to discuss the implications for bias. Three key heuristics – availability, representative, anchoring – are typically used when assessing probabilities.

The availability heuristic refers to a person's judgment being influenced by how easily historical events are recalled. As an example, an expert may believe the likelihood of a particular system failing is high because there was a recent failure event. This can lead to a biased assessment as the expert is not considering a full failure history of the system.

The representative heuristic refers to the similarity between an outcome and a population. For example, on assessing the probability that a new component design will have a mean lifetime in excess of 1000 operating hours, one may consider how well the existing component population represents the design being evaluated. Care is required using such a heuristic because an expert may assign an intuitive measure of similarity rather than considering how the probability will change with respect to that of the existing population.

The anchoring and adjusting heuristic refers to the situation when an expert provides an initial assessment and all future assessments are derived from an adjustments of the initial assessment. For example, consider the situation where a prior distribution for the mean time to failure for a component is required from an expert. An initial assessment may correspond to a best guess, from which the expert may anchor upon the initial assessment to adjust for upper and lower bounds from which the distribution is determined. Research shows that the adjustments made are typically insufficient (Winkler 1967, Alpert and Raiffa 1982).

4 ASSESSING PRIOR DISTRIBUTIONS

If an expert provides subjective probabilities that are coherent then these are not incorrect insofar as they reflect the beliefs of the expert. However, it is clear that different experts give different assessments, and so this raises the question of how one might assess “quality” or “usefulness” of

such distributions. As Cooke (1991) pointed out, there is no absolute measure of the correctness of a subjective assessment. However two measures are often used either quantitatively or qualitatively: calibration and information.

To assess the degree to which an expert's assessments are well calibrated requires the subjective probabilities to be compared with a collection of observed realisations. For example, two experts could assess the same event with one expert providing a probability of 0.9 and the other a probability of 0.2. This event will be realised or not. Since neither expert assigned a probability of 1 or 0, regardless of the outcome neither will be incorrect. The first expert would be calibrated if 90% of all events assessed by that expert at the 90% level are realised and 10% are not. While for the second expert 20% of all events assessed by that expert at the 20% level are realised and 80% are not. For this example, both experts could be perfectly calibrated but provide very different probabilities.

Direct assessment of calibration is only possible if we are able to make observations directly on the quantity of interest. In many cases, for example the uncertainty of a failure rate, this is not the case and so the distribution of any directly observable quantities (such as the number of units failing up to a given time) would be a predictive distribution obtained as a mixture of the sampling distribution weighted by the prior. Hence calibration measurements would naturally confound the sampling distribution and the prior.

A second consideration for the usefulness of an expert's prior distribution is information, which can be assessed through mean squared error between the observed realisations and the subjective probabilities, information scores, or a similar such formula. Stronger conclusions can be drawn from an informative prior distribution than from an uninformative one. Hence ideally would like to have an informative, well calibrated expert. Note that it is possible to have an uninformative calibrated expert. For example, if upon assessing a mean time to failure an expert

provides the entire range of positive real numbers below the median fifty percent of the time and above the median fifty percent of the time, then the expert is calibrated but not informative.

5 IMPROVING PROBABILITY ASSESSORS

The quality of subjective probabilities from an expert is dependent upon both the expert's experience and the method of elicitation. If the expert lacks experience, the prior distribution will be uninformative or misleading, regardless of the elicitation approach employed. Poorly designed elicitation techniques may degrade the quality of information provided by experts. Fischhoff (1989) proposes four necessary conditions to support improved judgement skills:

1. Abundant practice with a set of reasonably homogeneous tasks;
2. Clear-cut criterion events for outcome feedback;
3. Task-specific reinforcement;
4. Explicit admission of the need for learning.

There is extensive evidence that these are often not achieved in practice (Ferrell 1994, Wright and Bolger 1992).

Standard techniques for eliciting prior distributions that reflect the uncertainty of the expert are reported in the literature. See, for example Cooke (1991), Meyer and Booker (2001). The choice of technique depends upon the quantity of interest, the selection of which has been given little attention in the literature.

One important consideration when selecting the quantity of interest is feedback to the expert. This is considered crucial for calibrating the expert and should be event specific (Fischhoff 1989, Wright and Bolger 1992, Ferrell 1994). In other words, the feedback must be with respect to assigning probabilities to particular classes of relevant events and not only feedback on the ability of the expert to assign probabilities to any situation. To increase the effectiveness of feedback in terms of learning, conditions that influence the event should re-

occur as often as possible (Fischhoff 1989, Kadane and Wolfson 1997). Therefore the factors on which the measure is conditioned should be as few and general as possible.

6 GENERIC PROCESSES

A necessary first step towards developing a method for eliciting expert judgement is to start with a general approach that possesses the required key features for any elicitation process and to tailor it to the desired situation.

Phillips (1999), Clemen and Reilly (2001), Garthwaite et al (2005), for example, all describe generic processes. Another, first developed by the Stanford Research Institute (SRI) for eliciting expert judgement is discussed in Spetzler et al (1975), Ferrell (1985) and Merkhofer (1987). The issues addressed within the SRI process are common considerations within the alternatives. There are seven stages to the SRI approach: motivate, structure, condition, encode, verify, aggregate and discretise.

Motivating the expert can be achieved by explaining the elicitation process and how the results will be used. This phase can be used to enhance the comprehension of an expert regarding the process, encourage an expert to provide an accurate assessment of his/her understanding and determine potential biases, such as motivational biases. These include expert bias where an expert becomes overconfident merely because he/she has been given the title 'expert' and management bias where an expert provides goals as opposed to judgement. For example, an expert states the aspiration that there will be no faults within this system by time of manufacture, rather than an assessment of his beliefs.

Structuring involves defining the specific event being considered to ensure there is no ambiguity in the questions and to explore how an expert thinks about the quantity for which probability judgement is to be elicited. The aim is to manage cognitive bias by simplifying the complex task of assigning probabilities by disaggregating the quantity of interest into more elementary variables (Armstrong et al 1975). However, the unpacking principle (O'Hagan et al

2006) may be the consequence. This refers to the situation where the more detailed the description of the event, the greater the likelihood assigned to it. For example, an expert may provide an assessment for the probability of a component failing and subsequently during the elicitation process provide probabilities associated with causes of failure that may result in a system probability exceeding the initial assessment.

Sometimes experts may be unclear about whether they are being asked for conditional events or conjoined events ($P(A|B)$ or $P(A \text{ and } B)$), as these can be difficult to distinguish in natural language. Judgements can be conditioned upon all relevant information necessary to ensure an expert's awareness has been stimulated and to manage biases such as anchoring and availability.

Encoding refers to an expert providing numerical expression that reflects his/her uncertainty regarding an outcome. There are a number of techniques available for encoding subjective probabilities. Common approaches range from direct estimation whereby a person is simply asked for the probability to inferring probabilities from elicited preferences with specified lotteries. A few devices have been used to support such elicitation, such as probability wheels or visual analogue scales. An interesting discussion of alternative methods is given in O'Hagan et al (2006).

Encoding a probability distribution is more challenging because additional information is required than for an individual probability. One could partition the distribution into discrete intervals and elicit probabilities for the value of interest belonging to each of the intervals. Evidence exists to suggest that people are better at estimating the median and mode rather than the mean; moreover people are poor at assessing or interpreting the variance (O'Hagan et al 2006).

A popular alternative encoding procedure for distributions is the fractile method (Cooke 1991), where the expert assesses the median value of their subjective probability distribution

along with, the (25th, 75th) and the (5th, 95th) percentiles. Once these values have been elicited a parametric distribution can be sought to maximise fit. The main drawback of this technique is that the elicited data often reflect a central bias (Seaver et al 1978). However, after percentiles of the prior distribution have been assessed, graphical techniques can be applied to enhance the quality of the prior distribution (Chaloner et al 1993).

Verification is required to ensure that an expert has provided a reflection of his/her true beliefs. If problems are encountered then the previous stages are to be repeated.

If multiple experts are assessing the same quantity of interest, then their individual probability distributions should be aggregated. There are two approaches to aggregation – mathematical and behavioural. The former implies the experts should not influence each others decisions (Ferrell 1985). The latter requires experts to share their judgement and reassess their distributions and includes techniques such as Delphi (Ferrell 1985) and Nominal Group Technique (Moore 1987). There are several mathematical approaches to aggregation most of which aim to evaluate a weighted average across the experts. See Cooke (1991) for a fuller discussion.

It is often necessary to treat continuous random variables as discrete during the encoding stage. Discretising refers to techniques for fitting continuous distributions to the elicited data while preserving important moments. This is accomplished by dividing the range of all possible values for the uncertain variable into intervals, selecting a representative point from each interval and assigning that point the probability that the actual value will fall within the corresponding interval. The moments can be preserved through Gaussian quadrature techniques (Miller and Rice 1983).

Two elicitation processes specific to reliability include PREDICT (Kerscher (1998, 2003) and REMM (Walls et al (2006), Walls and Quigley (2001) and Hodge et al (2001)). Both provide modelling frameworks designed to estimate reliability throughout system development

beginning with a problem structuring phase, which consists of eliciting a graphical representation of the relationship between the potential failure modes and the system reliability. The graphs form the basis for the elicitation.

7 SUMMARY AND CONCLUSIONS

An overview of the key issues associated with the elicitation of a prior distribution has been presented; in particular, the shortcomings due to the prevalence of heuristics employed by an expert when assessing probabilities have been highlighted. Evidence indicates that an expert can improve his/her assessments given feedback about meaningful quantities of interest and this learning can be accelerated under the right conditions. Moreover, good subjective probability assessment requires a supportive process for the experts, whereby key stages are sequenced for managing specific biases. For further details about these issues see O'Hagan et al (2006), Meyer and Booker (2001) and Cooke (1991).

Practically, the design and implementation of an elicitation process requires considered project management to ensure that, for example, the correct experts are selected, the choice of elicitation method is relevant to the model requirements and parameterisation, the integrity of the probabilities assessed are maintained and that sufficient resources are obtained to support an effective, and efficient, process. See, Meyer and Booker (2001) for further discussion.

Theoretically, there is a need to infer from the assessments which probability distributions on model parameters are consistent. This approach has been developed by Cooke (1994) with more algorithms and underlying theory for probabilistic inversion in Kraan and Bedford (2005). Taking a more standard Bayesian perspective, Percy (2002), discusses the indirect assessment of parameters from generic prior distributions through the direct assessment of quantiles of the predictive distribution which will have observable outcomes. Gutierrez-Pulido et al (2005) take a similar line, considering both moments and quantiles of

the time-to-failure for a system as sources of information from which prior distributions can be fitted. Such methods could also be applicable to other Bayesian contexts where prior distributions on lifetime distribution parameters are to be assessed – for example in Bayesian accelerated or proportional hazards life modelling (Bunea and Mazzuchi 2005).

Elicitation of a prior distribution places a heavy cognitive burden on the subjective probability assessments by an expert. Yet observed data can exist for earlier generations of a system, component or process, although it is not clear how these data can be adapted to inform the assessment of a new design. Empirical Bayes provides a framework for integrating relevant historical data and hence can reduce the cognitive burden on an expert. See for example Quigley et al (to appear) and Bedford et al (2006).

REFERENCES

1. Alpert M and Raiffa H A progress report on the training of probability assessors. In D Kahneman, P Slovic and A Tversky (Eds) *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, 1982.
2. Armstrong, JS Denniston, WB and Gordon, MM The Use of Decomposition Principle in Making Judgements, *Oper. Behaviour Human Performance*, 1975 14: 257-263.
3. Bedford, T. and Cooke, R. *Probabilistic Risk Analysis: Foundations and Methods* Cambridge University Press, Cambridge, 2001.
4. Bedford T, Quigley J and Walls L Expert Elicitation for Reliable System Design, *to appear in Statistical Science*.
5. Bedford T, Quigley J and Walls L ‘Fault Tree Inference Through Combining Bayes and Empirical Bayes Methods’ *Proceedings of the ESREL Conference*, Estoril, Portugal 2006: pp 859-865.
6. Bunea, C. and Mazzuchi, T. A. Bayesian accelerated life testing under competing failure modes. In *Annual Reliability and Maintainability Symposium*, 2005

- Proceedings. Proceedings Annual Reliability and Maintainability Symposium. 2005:152–157.
7. Chaloner, K Church, T, Lois, T and Mattis, J Graphical Elicitation of a Prior Distribution for a Clinical Trial, *The Statistician*, 1993 42: 341- 353.
 8. Clemen RT and Reilly T Making Hard Decisions with Decision Tools, Duxbury Press 2001.
 9. Cooke, R. M. Experts in Uncertainty. Oxford University Press 1991.
 10. Cooke, R. M. Parameter fitting for uncertain models: modelling uncertainty in small models. *Reliability Engineering and System Safety* 44, 1994: 89–102.
 11. Evans, RA Bayes is for the Birds, *IEEE Transactions in Reliability*, 1989 38: 401.
 12. Ferrell, W Combining Individual Judgements, *Behaviour Decision Making*, G. Wright (ed.) Plenum Press 1985.
 13. Ferrell, W., “Discrete subjective probabilities and decision analysis: elicitation, calibration and combination”, in *Subjective Probability*, G.Wright and P. Ayton (ed.), 1994; John Wiley and Sons.
 14. de Finetti, B Theory of Probability, New York, Wiley 1974.
 15. Fischhoff, B Eliciting Knowledge for Analytical Representation, *IEEE Transactions on Systems, Man, and Cybernetics*, 1989 19: 448-61.
 16. Gartwaithe PH, Kadane JB and O’Hagan A Statistical Methods for Eliciting Probability Distributions, *JASA*, 100, 2005: 680-701.
 17. Gutierrez-Pulido, H., Aguirre-Torres, V., and Christen, J. A. A practical method for obtaining prior distributions in reliability. *IEEE Transactions on Reliability* 2005 54, 2:262–9.

18. Hodge R, Evans M, Marshall J, Quigley J and Walls L Eliciting engineering knowledge about reliability during design - Lessons learnt from implementation *Quality and Reliability Engineering International* 2001 17: 169-179
19. Kadane, J and Wolfson, L Experiences in Elicitation , *The Statistician*, 1997 47: 3-20.
20. Kahneman, D Slovic, P and Tversky, A Judgement Under Uncertainty: Heuristics and Biases. Cambridge University Press 1982.
21. Keeney, R.L. and von Winterfeldt, D., Eliciting probabilities from experts in complex technical problems, *IEEE Transactions on Engineering Management*, 1991 38: 191-201.
22. Kerscher, W. J., I., Booker, J. M., Bement, T. R., and Meyer, M. A. Characterizing reliability in a product/process design-assurance program. Annual Reliability and Maintainability Symposium 1998 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.98CH36161). IEEE, Anaheim, CA, USA, 1998:105–12.
23. Kerscher, W., Booker, J., and Meyer, M. Predict: A case study, using fuzzy logic. In Proceedings of the Reliability and Maintainability Symposium 2003.
24. Kraan, B. and Bedford, T. Probabilistic inversion of expert judgments in the quantification of model uncertainty. *Management Science* 2005 51, 6: 995–1006.
25. Merkhofer, M Quantifying Judgemental Uncertainty: Methodology, Experiences, and Insights *IEEE Trans. On Systems, Man. And Cybernetics*, 1987 17: 741 - 752.
26. Meyer, MA and Booker, JD, Eliciting and Analysing Expert Judgement: A Practical Guide, 2nd ed. ASA-SIAM Series on Statistics and Applied Probability 7 2001.
27. Miller, AC and Rice, TR Discrete Approximations of Probability Distributions, *Management Science*, 1983 29: 352-362.
28. Moore, CM Group Techniques for Idea Building, Sage 1987.

29. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T Uncertain Judgements Eliciting Experts' Probabilities, Wiley 2006.
30. Percy, D. F. Bayesian enhanced strategic decision making for reliability *European Journal of Operational Research* 2002 139, 1: 133–145.
31. Phillips LD Group Elicitation of Probability Distributions: Are Many Heads Better Than One? In J Shanteau, Barbara Mellors and David Schum (Eds) *Decision Science and Technology: Reflections on the Contributions of Ward Edwards*, Kluwer Academic Publishers 1999
32. Quigley J, Bedford T and Walls L Estimating Rate of Occurrence of Rare Events with Empirical Bayes: A Railway Application *Reliability Engineering and System Safety*.
33. Seaver, D, von Winterfeldt, D and Edwards, W Eliciting Subjective Probability Distributions on Continuous Variables, *Organizational Behaviour Human Performance*, 1978 21, 379-91.
34. Spetzler, C and Stael von Holstein, CA (1975) Probability Encoding in Decision Analysis, *TIMS*, **22**, 340 - 358.
35. Sunstein CR *Laws of Fear* Cambridge University Press 2005.
36. Tversky, A and Kahneman, D Judgement Under Uncertainty: Heuristics and Biases, *Science*, 1974 185: 257-263.
37. Walls L and Quigley J Building Prior Distributions to Support Bayesian Reliability Growth Modelling Using Expert Judgement *Reliability Engineering and System Safety* 2001 74: 117-128
38. Walls L, Quigley J and Marshall J Modeling to Support Reliability Enhancement during Product Development with Applications in the UK Aerospace Industry, *IEEE Transactions on Engineering Management* 2006 53 (2): 263-274.

39. Winkler, R The Assessment of Prior Distribution in Bayesian Analysis, *JASA*, 1967 62:
776-800.
40. Wright, G and Bolger, F (eds.) Expertise and Decision Support, Plenum 1992.