1-2005

Statistical Methods for Eliciting Probability Distributions

Paul H. Garthwaite
Open University, Milton Keynes

Joseph B. Kadane
Carnegie Mellon University, kadane@stat.cmu.edu

Anthony O'Hagan
University of Sheffield

Statistical Methods for Eliciting Probability Distributions

Paul H. Garthwaite, Joseph B. Kadane and Anthony O'Hagan*

January 4, 2005

Abstract

Elicitation is a key task for subjectivist Bayesians. While skeptics hold that it cannot (or perhaps should not) be done, in practice it brings statisticians closer to their clients and subject-matter-expert colleagues. This paper reviews the state-of-the-art, reflecting the experience of statisticians informed by the fruits of a long line of psychological research into how people represent uncertain information cognitively, and how they respond to questions about that information. In a discussion of the elicitation process, the first issue to address is what it means for an elicitation to be successful, i.e. what criteria should be employed? Our answer is that a successful elicitation faithfully represents the opinion of the person being elicited. It is not necessarily "true" in some objectivistic sense, and cannot be judged that way. We see elicitation as simply part of the process of statistical modeling. Indeed in a hierarchical model it is ambiguous at which point the likelihood ends and the prior begins. Thus the same kinds of judgment that inform statistical modeling in general also inform elicitation of prior distributions.

*Paul H. Garthwaite is Professor of Statistics, Faculty of Mathematics and Computing, The Open University, Milton Keynes, UK (email: P.H.Garthwaite@open.ac.uk); Joseph B. Kadane is Leonard J. Savage University Professor of Statistics and Social Science, Carnegie Mellon University, Pittsburgh, PA 15213 (email: kadane@stat.cmu.edu) and Anthony O'Hagan is Professor of Statistics at The University of Sheffield, Sheffield, UK (email: a.ohagan@sheffield.ac.uk). Professor Kadane's work was partially supported by NIH Grant IR01 GM868950-01 and by the National Science Foundation under Grant No. 0139911. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Professors Garthwaite and O'Hagan's work was supported by research funding from the UK department of Health and the Centre for the Bayesian Statistics in Health Economics, University of Sheffield.

The psychological literature suggests that people are prone to certain heuristics and biases in how they respond to situations involving uncertainty. As a result, some of the ways of asking questions about uncertain quantities are preferable to others, and appear to be more reliable. However data are lacking on exactly how well the various methods work, because it is unclear, other than by asking using an elicitation method, just what the person believes. Consequently one is reduced to indirect means of assessing elicitation methods.

The tool-chest of methods is growing. Historically the first methods involved choosing hyperparameters using conjugate prior families, at a time when these were the only families for which posterior distributions could be computed. Modern computational methods such as Markov Chain Monte Carlo have freed elicitation from this constraint. As a result there are now both parametric and non-parametric methods available for low-dimensional problems. High dimensional problems are probably best thought of as lacking another hierarchical level, which has the effect of reducing the as-yet-unelicited parameter space.

Special considerations apply to the elicitation of group opinions. Informal methods, such as Delphi, encourage the participants to discuss the issue in the hope of reaching consensus. Formal methods, such as weighted averages or logarithmic opinion pools, each have mathematical characteristics that are uncomfortable. Finally, there is the question of what a group opinion even means, since it is not necessarily the opinion of any participant.

Keywords: Bayesian, group decisions, heuristics and biases, prior distributions, subjective probability

1. THE ELICITATION CONTEXT

1.1. Introduction

Elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a (joint) probability distribution for those quantities. In the context of Bayesian

statistical analysis, it arises most usually as a method for specifying the prior distribution for one or more unknown parameters of a statistical model. In this context, the prior distribution will be combined with the likelihood through Bayes' theorem to derive the posterior distribution. However, this is not the only context in which elicitation is important.

Much of the literature on elicitation has been concerned with formulating a probability distribution for uncertain quantities when there is no data with which to augment the knowledge expressed in that distribution. This situation arises in decision making, where uncertainty about the 'state of nature' needs to be expressed as a probability distribution in order to derive (and then maximise) expected utility. Similarly, it arises in the use of mechanistic models. Such models are built in almost all areas of science and technology, to describe, understand and predict the behaviour of complex physical processes. The user is required to specify the values of appropriate model inputs, in order to run the model and obtain outputs, but there is generally uncertainty about the 'true' values of the inputs. It is then important to formulate that uncertainty and to propagate it through the model so as to quantify the uncertainty in model outputs.

It is convenient to think of the elicitation task as involving a *facilitator* who assists the *expert* to formulate the expert's knowledge in probabilistic form. In the context of eliciting a prior distribution for a Bayesian analysis, it is the expert's prior knowledge that is being elicited, but in general the objective is to express the expert's *current* knowledge in probabilistic form. If the expert is a statistician, or very familiar with statistical concepts, then there may be no formal need for a facilitator, but this is rare in practice. We shall see that elicitation is a complex process that demands a range of skills if it is to be done well, and the role of facilitator is an important one.

What does it mean for an elicitation to be done well? It is important to distinguish between the quality of an expert's knowledge, and the accuracy with which that knowledge is translated into probabilistic form. An elicitation is done well if the distribution that is derived accurately

represents the expert's knowledge, regardless of how good that knowledge is. The expert might, for instance, believe very strongly in a certain scientific hypothesis. Then the elicitation is accurate if it derives a suitably high probability for that hypothesis being true, even if it is subsequently found to be false. Even if the rest of the scientific community is much more sceptical, and inclined to give the hypothesis a low probability, this expert believes in the hypothesis and therefore accurate elicitation of this expert's knowledge and beliefs should derive a high probability for it.

To achieve accurate elicitation is by no means straightforward, even if we wish to elicit the expert's beliefs about just a single event or hypothesis (or equivalently, for a binary random variable). In this case, we require only a single probability, but the expert may be unfamiliar with the meaning of probabilities. Even when the expert is familiar with probabilities and their meaning, it is not easy to assess a probability value for an event accurately.

If we now consider the task of eliciting a distribution for a continuous random variable X, then implicitly this involves eliciting an infinite collection of probabilities $F(x) = P(X \leq x)$ for all the possible values of x. This is clearly impossible, and in practice an expert can only make a finite number (and usually a rather small number) of statements of belief about X. These might take the form of individual probabilities or quantiles of the distribution, i.e. $P(X \leq x)$ for a few distinct values of x, or might be some other summaries of the distribution such as a mode. When it comes to a joint distribution for a collection of random quantities, the magnitude of the elicitation task is very much larger still.

Given the difficulty involved, why is it worth the effort to attempt elicitation? One reason has to do with the use of elicitations to make decisions. Often a reasonable goal for elicitation is to capture the "big message" in the expert's opinion. The details, for example the exact shape of the expert's opinion, may not matter for the decision to be reached. Even when the decision is sensitive to the exact shape of the elicited distribution, it is not the decision, but rather the expected utility

of the decision, that matters. And expected utility of the optimal decision is very often robust to details of the expert's opinion.

A second reason why elicitation is worthwhile has to do with the use of elicitations to make inferences, and in particular for making possible the calculation of posterior distributions. In such a situation, elicitation encourages the expert and the facilitator to consider the meaning of the parameters being elicited. This has two helpful consequences. First, it brings the analysis closer to the application by demanding attention to what is being modelled, and what is reasonable to believe about it. Second, it helps to make the posterior distributions, once calculated, into meaningful quantities.

Elicitation is properly conceived of as part of the familiar process of statistical modelling. Statisticians are used to stating a likelihood for an applied problem. This is an opinion about how the data are generated, conditional on certain parameters. Hierarchical models, such as random effects models and models with latent variables, involve distributions on some parameters, conditional on yet others. What we are calling elicitation in this article is merely the final step in this process, the statement of probability distributions of the highest level parameters in such a hierarchy. It is well to keep in mind that the usual principles of statistical modelling apply to elicitation as well.

1.2. The elicitation process

Figure 1 is a schematic representation of the elicitation process in terms of four separate stages.

1. The *setup* stage consists of preparing for the elicitation — selecting the expert(s), training the expert(s), identifying what aspects of the problem to elicit, etc.

2. We then *elicit* specific summaries of the experts' distributions for those aspects. This is obviously the core of the process, and one where psychologists have contributed at least as

much to the methodology as statisticians.

3. The next stage is to *fit* a (joint) probability distribution to those summaries. In practice, this stage often blurs with the previous one, in the sense that the choice of what summaries to elicit is often influenced by choice of what distributional form the facilitator intends to fit.

4. Elicitation is almost invariably an iterative process, and the fourth stage involves assessing the *adequacy* of the elicitation, with the option then of returning to the second stage and eliciting more summaries from the expert(s).

This chapter is structured in accordance with this schematic. The remainder of Section 1 concerns topics relevant to the setup of the elicitation; Section 2 deals with the interaction with the expert to elicit specific summaries; Section 3 addresses how to fit a probability distribution to the elicited summaries; Section 4 deals with assessing the accuracy of elicitation. Questions which arise when beliefs are elicited from several experts are considered in Section 5, and a final Section 6 offers some discussion and challenges for future research.

1.3. Whose beliefs?

We have presented elicitation as the process of formulating in probabilistic terms the beliefs of an expert, but who is the expert? Use of the term 'expert' suggests an emphasis on persons to whom society and/or his/her peers attribute special knowledge about the matters being elicited. In practice, we often seek to identify the best available knowledge about the quantities of interest, and for this purpose we can regard the 'expert' as a real expert. There are other uses of elicitation, however, in which the expert has little or no expertise in the usual sense of that word. For example, to study adolescent decision making around risky behaviours, one might want to ask adolescents how they perceive those risks. Here the very point of the study is the lack of expertise of the

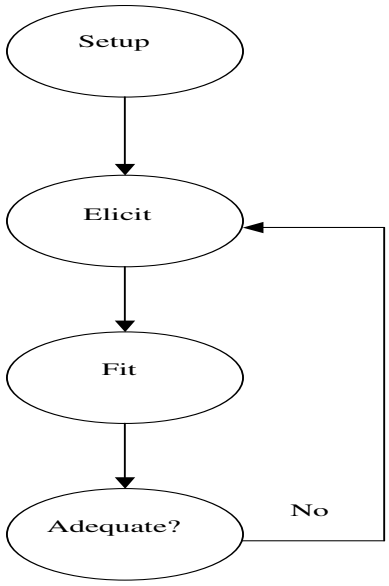

Figure 1: The elicitation process.

'experts'.

The simple answer to the question "who is the expert?" is that the expert is the person whose knowledge we wish to elicit; the term 'expert' does not necessarily signify any more than that.

An important point to bear in mind when eliciting from an acknowledged expert is that expertise can bring biases if the expert has some kind of personal interest in the result. Suppose, for example, that a radiation expert is asked for an opinion about how serious a health problem is engendered by the radiation release at Chernobyl. Such an expert may have spent much of his or her adult life becoming an expert on radiation. How well that expertise pays off in terms of social attention (and grants) depends on how urgent society perceives the issues the expert studies to be. Hence such an expert has an incentive to emphasize the dangers. For more on this kind of bias, see Kadane and Winkler (1988).

In Section 5 we consider the case of multiple experts, where often the desire is to combine the expertise of several people. Then it is sensible to try to ensure that the experts' knowledge is

complementary. Where their knowledge overlaps, it is more difficult to account for this, as we shall see in Section 5, and there is less gain from using the extra experts.

1.4. Conducting the elicitation

We outline here various aspects of good practice in the conduct of elicitations. Many of these can be ignored in an informal elicitation, but they are important considerations wherever substantive decisions or inferences may depend on the expert's distribution.

- The objective is to elicit a distribution to represent the expert's current knowledge. It is very useful to have a summary of what that knowledge is based upon.
- Any financial or personal interest that the expert might have, in the inferences or decisions that will depend (even marginally) on the expert's distribution, must be declared.
- Training should be given to familiarise the expert with the interpretation of probability and with whatever concepts and properties of probability will be required in the elicitation. It is useful to run through a dummy elicitation exercise to provide practice in the protocol that the facilitator proposes to use.
- A record should be kept of the elicitation. This should ideally set out all the questions that were asked by the facilitator together with the expert's responses, as well as the process by which a probability distribution was fitted to those responses.

Note that these recommendations apply whatever specific protocol the facilitator will use to elicit the expert's beliefs. We now proceed to detailed discussion of how to construct a suitable protocol.

2. Psychological Considerations and Eliciting Summaries

An elicitation method forms a bridge between an expert's opinions and an expression of these opinions in a statistically useful form. Thus, the development of an elicitation method requires some understanding of both the psychological part of the bridge and the statistical part. As Hogarth (1975, p. 284) points out "*... assessment techniques should be designed both to be compatible with man's abilities and to counteract his deficiencies*". In this section we present some results from psychological research that should be taken into account when forming methods of quantifying an expert's opinion. Much of the fundamental work in this area stems from the 1960s and early 1970s, and good reviews of this research are given in Hampton et al (1973), Hogarth (1975), Huber (1974), Lichtenstein et al (1982), Peterson and Beach (1967), Slovic and Lichtenstein (1971) and Tversky (1974). Later reviews are given in Chaloner (1996), Cooke (1991), Hogarth (1987), Kadane and Wolfson (1998), Meyer and Booker (2001), Morgan and Henrion (1990), Wallsten and Budescu (1983) and Wolfson (1995).

2.1. Heuristics and biases

A body of psychological research has been concerned with the question of how a person assesses the probability of an event, or how he judges which of two or more events is the more likely to occur. It appears that intuitive judgements in these tasks are based on a limited number of mental operations, or heuristics. In general these heuristics are quite effective, but they can lead to severe errors and systematic bias. Elicitation techniques commonly require novel assessment tasks. An appreciation of the strategies people use to quantify their opinions can give an indication of how (and how well) these tasks might be performed (Meyer and Booker, 2001).

One commonly used heuristic is *judgement by representativeness.* This is applicable for questions of the form: What is the probability that an object A belongs to a class B? What is the probability

that event A will generate an event B? In answering these questions, which in effect require the probability $P(B \mid A)$ to be assessed, people typically compare the main features of A and B and assign a probability depending on the degree of similarity between them. A common error made with this kind of judgement is that little or no attention is paid to the unconditional probability of B. As an illustration, consider the problem of evaluating the probabilities that an individual, Mr. X, who has been described as "meticulous, introverted, meek and solemn" is engaged in one of the following occupations: farmer, salesman, pilot, librarian, physician. Most people perform this task by assessing the similarity of Mr. X to the stereotype of that occupation, and order the occupations by the extent to which Mr. X is representative of these stereotypes. They completely ignore base rates, such as the relative number of salesmen to librarians, and assign a high probability to Mr. X being a librarian (Kahneman and Tversky, 1973). Similar results have been obtained by Hammerton (1975), and Nisbett et al. (1976).

Another commonly used heuristic is *judgement by availability.* This is used when a person estimates the frequency of a class or the probability of an event by the ease with which examples are recalled or occurrences come to mind. Examples from large classes are usually recalled better and faster than examples from less frequent classes, and likely occurrences are easier to imagine than unlikely ones. Hence, mental availability is often a helpful indicator for the assessment of frequency and probability, but availability is also affected by factors other than frequency or probability. For example, suppose you are asked whether a randomly chosen word from an English text is more likely to start with an "r" or have "r" as its third letter. It is easier to recall words by their starting letter (e.g. *r*ed, *r*ank, *r*ogue, *r*oad, *r*ope, ...) than by their third letter (e.g. pa*r*k, bi*r*d, wi*r*e, ...). Hence, most people judge that "r" is more likely to be the first letter of a word, rather than the third letter, although the reverse is true (Tversky and Kahneman, 1973). Recall is also affected by factors such as familiarity, salience and recency, and newsworthy events also

impact disproportionately on our memory, so you might overestimate the probability of a plane crash with fatalities, for example, particularly if such a crash has happened recently. Thus the judgement-by-availability heuristic, though useful, can lead to marked error.

Perhaps the heuristic most widely used for probability assessment is *judgement by anchoring and adjustment.* With this strategy, a person estimates an unknown quantity by starting from some initial value and then adjusting it to obtain a final estimate. The starting value, which is usually termed the *anchor*, could be suggested by the nature of the problem or the way it is formulated. Regardless of the source of the starting value, the adjustment is usually too small (Slovic, 1972), a phenomenon called *anchoring.* An experiment conducted by Tversky and Kahneman (1974) elegantly demonstrated this effect. Subjects were asked to estimate various quantities, stated in percentages (e.g. the percentage of African countries in the United Nations). They were given randomly chosen starting values and had first to decide whether the value they had been given was too high or too low, and then adjust it until they reached their best estimate. Through insufficient adjustment, subjects whose starting values were high ended up with substantially higher estimates than those who started with low values. For example, the median estimates of the percentage of African countries in the U.N. was 25% for subjects who received 10% as their starting point and 45% for those who received 65% as their starting point.

As mentioned above, when subjects use the heuristic judgement-by-representativeness to assess probabilities, they tend to ignore prior probabilities. However, many experiments have shown that if subjects are made aware of the prior probabilities and are asked to modify these in the light of fresh sample data, then the assessed posterior probabilities are too close to the prior probabilities, compared with the revision indicated by Bayes' theorem. i.e. the subjects' modifications to the prior probabilities are insufficient. This type of insufficiency when modifying probabilities to reflect new data is referred to as *conservatism* (Edwards and Phillips, 1964). One possible explanation of

this phenomenon is that subjects use the anchoring and adjustment strategy; the prior probability acts as the anchor and the adjustment is insufficient.

A typical experiment to demonstrate conservatism is one in which subjects assess the probability that coloured poker chips have been drawn from one of two bookbags, where the two bags contain different compositions of chips e.g. 70% red and 30% blue in one bag, and 30% red and 70% blue in the other. A coin is tossed to select one of the bags, so the prior probability of each bag is 0.50, and the experimenter then draws a succession of chips with replacement from the selected bag, indicting their colour to the subject. Having observed the sample evidence, the subject states his posterior probability that the bookbag sampled contained the predominantly red or the predominantly blue proportion of chips. Subjects' revisions are normally conservative compared with the objective probability calculated by Bayes' theorem. For example, when the sample contains eight red and four blue chips, subjects commonly give a probability for the "red" bookbag of about 0.75, whereas the posterior probability calculated by Bayes' theorem is 0.97.

Several studies have attempted to counteract conservatism through varying the experimental procedure: in case subjects avoided approaching the bounds of the probability scale, unbounded odds estimates have been used instead of probability estimates; rewards have been added to provide an incentive to perform well; sample sizes, sequence lengths and prior probabilities have been varied. Some of these changes have influenced the degree of conservatism, but they have not eliminated it. (See e.g. Peterson and Beach, 1967, pp 32-33 for a review of such experiments.) In some experiments though, the basic experimental situation has been modified so as to make it more complex and, in these more complex situations, conservatism is not always a dominating influence (Youssef and Peterson, 1973).

Research has also demonstrated that, in many circumstances, people expect a sample from a population to represent all the essential characteristics of that population, even if the sample is

small. Tversky and Kahneman (1971) refer to this false logic as the 'law of small numbers', which asserts that the law of large numbers applies to small numbers as well. One experiment which showed this fallacious belief used, as subjects, audiences at two psychology meetings. Through a questionnaire, the psychologists were asked to decide sample sizes and also to relate sample sizes to inferences they would make from hypothesis tests. The following is a typical example of the questions asked:

"Suppose you have run an experiment on twenty subjects and have obtained a significant result which confirms your theory ($z = 2.23$, $p < 0.05$, two tailed). You now have cause to run an additional group of ten subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?" (Tversky and Kahneman, 1971, p105)

The median answer from the two groups was 0.85. However, if one assumes a non-informative prior distribution for the mean before the first sample was taken, then the true probability is only 0.48. The error is readily attributable to belief in the 'law of small numbers'; people expect all samples to have virtually identical characteristics. For similar reasons, answers to other parts of the questionnaire indicated that the majority of respondents (a) were too easily convinced by early results from a small experiment, (b) tested their research hypotheses on small samples without realising the high odds against detecting the effects being studied and (c) rarely attributed unexpected results to sampling variability because they found a causal explanation for every observed effect.

The results demonstrating the 'law of small numbers' and those demonstrating conservatism appear somewhat contradictory, the former suggesting that people overestimate the value of sample evidence while the latter suggest that people underestimate it. One conjecture is that conservatism has an effect if people first formulate their opinion prior to being given the sample evidence, while the 'law of small numbers' has an effect if people obtain the sample evidence before first formulating their opinion (Garthwaite, 1983, p 17).

Another form of systematic error that affects people's judgements is called *hindsight bias*. This can arise when people are asked to assess their *a priori* probability of an event that has actually occurred. For example, they might be asked whether the dismantling of the Berlin wall was predictable before it happened: "In 1988, what was the probability that the Berlin wall would come down within the next five years?" In 1988 it may have seemed unlikely that communism would soon collapse and East and West Germany re-unite, but *with hindsight* the economic problems of communist countries make it seem almost inevitable. Knowledge of what has occurred tends to distort memory and people tend to exaggerate their *a priori* probability for an event that has occurred. An experiment that shows this clearly was conducted by Fischhoff and Beyth (1975). Just before President Nixon's visit to China and the USSR in 1972, subjects were asked to assess the probabilities of various possible outcomes of his visit, such as "President Nixon will meet Mao at least once" and "The USA and the USSR will agree to a joint space program". Shortly after the visit and without forewarning, the subjects were asked to recall the probabilities they had given to these events before Nixon's visit, and they were also asked which events they thought had actually occurred. Results showed that subjects generally overestimated their *a priori* probabilities for the events they thought had occurred, and underestimated their *a priori* probabilities for the events they thought had not occurred.

We have described only some of the heuristics that people employ to make numeric judgements, and biases that affect such judgements. They were chosen for their relevance to elicitation methods. For example, conservatism, hindsight bias and the 'law of small numbers' relate to people's response to data, and a variety of elicitation methods use the impact of sample data on opinion to quantify beliefs about variances and covariances, as will be seen in Section 1.3. An extensive list of heuristics and biases in human judgement is given in Hogarth (1987, pp 216–222).

2.2. What summaries to elicit

In designing an elicitation method there is usually choice as to which quantities the expert is asked to assess and, if possible, quantities should be chosen that are usually assessed reasonably competently. People's ability to estimate simple statistical quantities, such as means and variances, has been examined in psychological research over several decades. Such quantities could constitute part of the elicitation method for almost any form of prior distribution.

Several experiments investigate subjects' capability at judging sample proportions (Erlick, 1964; Nash, 1964; Pitz, 1965, 1966; Shuford, 1961; Simpson and Voss, 1961; Stevens and Galanter, 1957). In these experiments, binary data were displayed to subjects for a limited period of time and they were then asked to estimate one of the sample proportions. For example, Shuford (1961) projected 20x20 matrices onto a screen, one at a time. The elements of each matrix were red squares and blue squares, and subjects observed a matrix for 1 second in some trials and for 10 seconds in others. After each trial, subjects had to estimate the proportion of squares that had been, say, red. In this and similar experiments, subjects generally assessed the sample proportion very accurately, with the mean of subjects' estimates differing from the true sample proportion by less than 0.05 in most cases.

Similar experiments have been used to investigate people's ability at estimating measures of central tendency (Beach and Swenson, 1966; Peterson and Miller, 1964; Spencer, 1961, 1963). Typically, a sample of numbers is displayed to subjects who are asked to estimate the mode, median or mean of the sample. When the sample distribution is approximately symmetric so that these three measures are numerically similar, subjects' estimates have shown a high degree of accuracy (Beach and Swenson, 1966; Spencer, 1961). However, an experiment conducted by Peterson and Miller (1964) used a sample drawn from a population whose distribution was highly skewed; subjects'

assessments of the median and mode were again reasonably accurate, but assessments of the mean were biased towards the median.

In most applications, to determine a prior distribution will require the variances of unknown scalar quantities and/or sample errors to be estimated. Regrettably, it seems that people are poor both at interpreting the meaning of 'variance' and at assigning numerical values to it. When estimating relative variability, empirical evidence indicates that people are influenced by the mean of the stimuli and estimate the coefficient of variation, rather than the variance. For example, Hofstatter (1939) obtained assessments of the variability in the lengths of sticks tied in bundles. He found that assessments increased with the sample variance, as it should, but as the means increased, the assessments decreased. Lathrop (1967) has replicated this latter result. Even allowing for the effect of means, systematic differences still arise between intuitive judgements of sample variance and the objective values. If large deviations from the mean predominate as when, for example, the sample is drawn from a population whose distribution is bimodal, then the variance is overestimated. On the other hand, if small deviations from the mean predominate as when, for example, the population distribution is normal, then the variance is underestimated (Beach and Scopp, 1967).

One way of eliciting variances that avoids their direct assessment is to elicit credible intervals; such intervals are useful in their own right and can yield estimates of variances if suitable distributional assumptions are made (cf Section 3.2). There are two main approaches to assessing credible intervals for a scalar quantity, the *fixed interval method* and the *variable interval method*. Let X denote the scalar quantity of interest. With the fixed interval method, the range of values that X can take is partitioned into intervals by the statistician/psychologist organizing the elicitation method. For each interval, the expert assesses the probability that X will fall in that interval. (In principle, assessed probabilities must sum to one.) The probabilities are also sometimes elicited

through odds assessment. The expert first indicates the interval which she believes is most likely to contain X. Then, for each of the other 'less likely intervals', she states odds that X will take a value in the 'less likely interval' as opposed to the 'most likely interval'. The probabilities associated with each interval are then calculated by imposing the constraint that their sum must be one.

With the variable interval method, the expert identifies points that correspond to specified percentiles of her subjective distribution. A method of bisection is often used, which entails a sequence of questions of the following form:

Q1. Can you determine a value (the expert's median) such that X is equally likely to be less than or greater than this point?

Q2. Suppose you were told that X is below your assessed median. Can you now determine a new value (the lower quartile) such that it is equally likely that X is less than or greater than this value?

Q3. Suppose you were told that X is above your assessed median. Can you now determine a new value (the upper quartile) such that it is equally likely that X is less than or greater than this value?

An advantage of this line of questioning is that only judgements of equal odds are required, an intuitively easier task than specifying percentiles that divide a probability in the ratio of, say, 4:1.

Many experiments have examined people's performance at assessing credible intervals. If the credible intervals calibrated well with reality, then the proportion of $p\,\%$ central credible intervals that contain the true value of X (the quantity to which they relate) should be about $p\,\%$. (The concept of calibration is discussed further in section 4.4). The experiments show that assessing credible intervals is a task that people perform reasonably well, but there is a clear tendency for central credible intervals to be too short, so that this proportion is less than $p\,\%$. This bias is referred

to as *overconfidence*; people believe they are more accurate in estimating X than is justified (Keren, 1991). Lichtenstein et al. (1982, pp. 325-326) tabulate 28 sets of results in which 50% central credible intervals were elicited for continuous scalar quantities. Less than 50% of the intervals contained the true value of the scalar in 23 of the 28 sets, and the proportion never exceeded 57%. The evidence is conflicting as to whether the fixed interval method or the variable interval method gives better calibration; Seaver et al (1978) found the fixed interval method performed better, while Murphy and Winkler (1974) found the converse. With the variable interval method, it is also unclear which percentiles should be elicited. The median and quartiles are most commonly assessed (using the method of bisection) and while this has sometimes given good results (e.g. Murphy and Winkler, 1974; Peterson et al, 1972), other empirical work has found that overconfidence is less if the 33 and 67 percentiles are assessed (Barclay and Peterson, 1973; Garthwaite and O'Hagan, 2000). Most empirical work has involved scalar quantities and the results may not generalize to more complex models. For example, Garthwaite (1989) elicited 50% central predictive intervals for the dependent variable in a simple linear regression model and found that far more than 50% of the intervals contained correct values. The subjects had drawn graphs to help make their assessments and this may have improved the accuracy of their median assessments and led to predictive intervals that were more likely to contain a true value.

Much empirical research has investigated subjects' ability to assess the extreme "tails" of a distribution. For example, Alpert and Raiffa (1969) used the variable interval method to elicit 98% central credible intervals. They asked "almanac" questions of the following kind:

How many foreign cars were imported into the U.S. in 1968?

(a) *Make a high estimate such that you feel there is only a 1 percent probability the true answer would exceed your estimate.*

(b) *Make a low estimate such that you feel there is only a 1 percent probability the true answer would be below this estimate.*

(Alpert and Raiffa, 1969, pp.16-17)

It should have been somewhat of a "surprise" to a subject to find the true value of a quantity falling outside an interval. 43% of all assessments produced such surprises. This information was given as feedback before a second session. In this second session 23% of the assessments produced surprises, which is still very high. One reason for the large number of surprises is that assessing tails of distributions is a difficult task, mainly because it requires the consideration of events that are unlikely, so that comparisons do not come readily to mind. (It is unfortunate that assessing probabilities for rare events is difficult, as expert opinion is of paramount importance when sample data is scarce.)

Task characteristics have an effect on the way an expert views a problem and the assessments that are elicited. Visual aids to help people quantify their opinions have been tried, such as urns full of coloured balls (Raiffa, 1968), light pens on coloured screens (Barclay and Randall, 1975), or simply asking assessors to mark a point on a line whose endpoints are 0 and 1 (for probabilities) or 0% to 100% for proportion. Probability wheels (Spetzler and Stael von Holstein, 1975) are another visual aid and they have been used with some success (Morgan and Henrion, 1990, p126). In its simplest form a probability wheel is a round pie-shaped disc of one colour that is partly covered by a 'slice' of a different colour, and a pointer. The size of the slice can be varied and the expert adjusts its size so that if the pointer is spun, then the probability it lands within the slice is equal to the expert's probability for some specified event.

Efforts to influence how people consider probabilities have also been explored, such as asking them to suggest scenarios that would lead to an unlikely event (Slovic and Fischhoff, 1977), influence

diagrams (Howard and Matheson, 1984), getting subjects to think carefully about the substantive details of each judgement (Koriat et al., 1980), and disaggregating an implicit hypothesis into its constituent hypotheses (Johnson et al., 1993). As an example of the effect of disaggregation, Fischhoff, Slovic and Lichtenstein (1978) questioned experts (car mechanics) about the probable reasons for a car not starting. The experts assessed the probability that it would not start "for some reason other than the battery, engine or fuel system" and their average probability was 0.22. They also assessed the probability that it would not start for more specific reasons: failure of the ignition system, failure of the starting system, etc. Combining the probabilities of the latter disaggregated reasons gave an average of 0.44 as the probability it would not start for reasons other than the battery, engine or fuel system. This result is consistent with other empirical work; there is ample evidence that the sum of separate probability assessments for constituent hypotheses generally gives a much larger probability than a single probability assessment of the combined hypothesis that they form (see Tversky and Koehler (1994) for a review). Morgan and Henrion (1990, p116) comment, "It has become something of an article of faith in the decision analysis community that disaggregation of an elicitation problem holds the potential for significantly improved performance on many assessment tasks".

There has been research aimed at converting probabilistic phrases (such as "quite likely" and "extremely probable") into numeric values (Wallsten et al., 1986; Mosteller and Yountz, 1990), and at differentiating situations where verbal expressions of probability are preferable to numeric ones, or vice-versa (Winschitl and Wells, 1996). People are generally more comfortable expressing their uncertainty in verbal terms rather than numerically. Unfortunately, there is considerable variation in the probabilities different people attach to the same phrase, and the context also affects the probability a person associates with a phrase (Lichtenstein and Newman, 1967; Beyth-Marom, 1982; Wallsten et al., 1986). The response mode in which subjects are asked to give assessments

also affects judgements. For example, Gigerenzer (1996) found that numeric expressions that are formally equivalent, such as frequencies and probabilities, are not always treated as equivalent in subjective uncertainty judgements. Also, it seems better to elicit probabilities in terms of populations of events, such as *What proportion of students starting a PhD will complete it within five years?* rather than as single (one-shot) events, like *If a new PhD student is picked at random, what is the probability that he or she will complete the PhD within five years?* (Gigerenzer, 1996; Koehler, 1996)

As noted earlier, a good elicitation method should yield a probability distribution that accurately reflects the expert's opinion, but this is hard to check and a pragmatic alternative is to compare assessed distributions with true values when these are known (see the discussion of calibration in Section 4.4). Several experiments have attempted to train subjects in order to improve the calibration/objective accuracy of their assessments. These have typically found that objective accuracy is improved substantially by training, but that biases such as overconfidence are tempered rather than eliminated (Schaefer and Borcherding, 1973; Lichtenstein and Fischhoff, 1980). In these experiments, training has usually taken the form of feedback; subjects are told the correct values after making the assessments and the trainer stresses the direction of biases and how the expert might reduce them. The benefits of effective feedback can be seen in the performances of weather forecasters. Weather forecasters make regular predictions for the same quantities each day (such as temperatures and the probability of precipitation), and thus soon learn the accuracy of their forecasts. Experiments have generally found them to be quite well-calibrated. For example, Peterson et al (1972) conducted an experiment with two meteorologists in which they gave forecasts of the maximum and minimum temperatures on the following day. Using the method of bisection, the meteorologists expressed their forecasts in terms of 50% central credible intervals. They showed good calibration: out of 55 forecasts, 28 fell inside the 50% credible intervals, 18 fell outside and 9

fell on the boundaries.

In the main, research concerned with descriptive statistics for scalars has produced relatively clear-cut conclusions. People are capable of estimating proportions, modes and medians of samples. We are slightly less proficient at assessing sample means if the sample distribution is highly skewed and we often have serious misconceptions about variances. We are reasonable at quantifying our opinions as credible intervals using the fixed interval and variable interval methods. However, there is a general tendency for the assessed distributions to imply a greater degree of confidence than is justifiable. Practice, coupled with feedback, will reduce this bias, but assessing the extreme tails of distributions is difficult (e.g. 98% credible intervals) and while training should reduce bias, it will not eradicate it. Visual aids can prove useful and task characteristics often have a marked impact on the assessments that are elicited.

2.3. Multivariate elicitation

When the expert's opinion is sought on two or more unknown variables, then the output of the elicitation should be the expert's *joint* probability distribution for those variables. The task is now more complex than when eliciting a distribution for a single variable, and the facilitator must inevitably ask more complex questions.

An important special case is where the variables are independent, meaning that if the expert were to obtain new information about some of the variables it would not change her beliefs about the others. The concept of independence is straightforward to explain, and independence between variables is a relatively simple judgement for the expert to make. It is also a very convenient judgement, because when all the variables are independent their joint distribution is just the product of their marginals. The elicitation exercise then reduces to eliciting the expert's beliefs about each variable separately, so only univariate elicitation techniques are required. Utilising independence to

decompose a multivariate elicitation task into simpler univariate tasks is consistent with the idea of disaggregation.

Psychological research also indicates that, when events are independent, joint probabilities should be assessed via univariate probabilities, as people exhibit systematic bias when making joint probability assessments. In particular, people tend to overestimate the probability of conjunctive events and underestimate the probability of disjunctive events. For example, Bar-Hillel (1973) found that people tended to overestimate the probability of drawing a red marble seven times in succession from a bag containing 90% red marbles and 10% white marbles, and underestimate the probability of drawing a red marble at least once in seven successive draws from a bag containing 10% red marbles and 90% white marbles. These errors can be explained as the result of anchoring: the probability of an elementary event provides an obvious starting point for estimating the probability of both conjunctive and disjunctive events. For conjunctive events, the probability of the elementary event must be reduced, which is done insufficiently, and for disjunctive events it must be increased, which is again done insufficiently.

Discussion of the physical or historical relationships among variables can make judgements of independence or conditional independence clear. With many elicitation methods it is transparent as to what assessments would correspond to independence and, in application of these methods, subjective independence between some pairs of parameters is often observed; see examples in Garthwaite and Dickey (1991, 1992). It should be noted though, that assumptions of independence often make assessment tasks easier for an expert. For example, if an expert has assessed the marginal distribution of X, and X and Y are independent, then the conditional distribution of $X \mid Y$ is easily specified as "no change". It may be that experts are too willing to accept independence where it does not strictly apply.

Even where variables are dependent, it may be possible to restructure the problem by expressing

it in terms of independent variables. An example might be where we seek a medical expert's opinion on the effectiveness of two treatments in a clinical trial. Letting X and Y denote the relevant measures of effectiveness of the two treatments, we would not typically have independence between X and Y. If the expert learned that X, the effectiveness of the first treatment, was higher than she originally expected, then this would generally lead her to have an increased expectation of the effectiveness Y of the second treatment. This may be because the expert believes that the treatments act in similar ways, but it may also be because of uncertainty in patient recruitment. That is, if X is smaller than expected, say, this may be because the trial has recruited patients who are more ill, and will thereby suggest a smaller value for Y. However, the expert might be willing to accept independence between two *functions* of X and Y. For instance, it may be reasonable to suppose independence between X and $Z = Y/X$. Here, Z is the relative effectiveness of treatment 2 over treatment 1. Such a structure is often appropriate where treatment 1 is standard care or placebo and treatment 2 is a new or active treatment. Where both treatments are new, the asymmetry of the preceding structure may not be appealing, but the expert might be happy to express independence between $(X+Y)/2$, the mean effectiveness, and the difference $Y-X$. Bayesian hierarchical models are natural examples of structuring dependent variables in terms of conditional independence. O'Hagan (1998) emphasises the role of structuring as an aid to elicitation. Kadane and Schum (1996) provide an extended example of complex structuring of beliefs.

Where variables are dependent and cannot obviously be reduced to independence in this way, we cannot escape from the complexity of multivariate elicitation. We can (and generally should) elicit summaries of the expert's marginal distributions, but these no longer characterise the joint distribution completely. The question then arises as to which summaries of the expert's joint distributions are most effective and reliable to elicit.

Although statisticians usually model dependence in terms of correlations, directly eliciting cor-

relation coefficients or covariances might be expected to encounter at least as many problems as directly eliciting means and variances in univariate elicitation. Psychological research has primarily considered eliciting correlation between variables that might be considered as drawn from some population. For example, Clemen et al (2000) examine the following six methods of eliciting a correlation between weight and height in a population of male MBA students.

1. Verbal (non-numeric) description of the strength of a correlation on a 7-point scale ranging from 'very strong negative relationship' to 'very strong relationship'. (Clemen et al made strong assumptions to convert the verbal assessments to correlations.)

2. Direct assessment of the correlation by specifying a value between -1 and 1.

3. Ask the subject to imagine that a person has been picked at random from the population. Give the person's percentile for one variable, and ask the subject to assess the person's percentile for the second variable.

4. Ask the subject to imagine that two people, A and B, have been picked at random from the population. Conditional on A being greater than B for one variable, ask the subject to assess the probability that A is also bigger than B for the other variable.

5. Ask the subject to imagine that a person has been picked at random from the population. The subject is asked to assess the probability that for both variables the person is below a specified percentile.

6. Ask the subject to imagine that a person has been picked at random from the population. Conditional on the person being below a specified percentile for one variable, ask the subject to assess the probability that the person is also below that percentile for the second variable.

Clemen et al found that method 2 performed best. This is surprising since others have suggested that the direct assessment of moments is a poor method of quantifying opinion (Morgan and Henrion, 1990; Kadane and Wolfson, 1998; Gokhale and Press, 1982). Method 4 asks for a concordance probability to be assessed, which can equated to a value of Kendall's τ. Assumptions of normality are then made so as to relate Kendall's τ to the Pearson correlation coefficient. Assessment of concordance probabilities to examine correlation has been examined by Gokhale and Press (1982), who found it preferable to alternative methods they consider, and by Kunda and Nisbett (1986), who concluded that reasonably accurate correlation estimates are obtained provided (a) subjects are very familiar with observations from the population in question, and (b) the data relate naturally to a numeric scale. In several experiments, subjects have been shown samples from a bivariate population and then asked to judge the 'degree of relatedness' between the two variables. In these experiments, it has been found that subjects make use of only a limited portion of the available data, sometimes basing their judgements on just the proportion of time the positive outcome for one of the binary variables occurred with a positive outcome for the other (Smedsland, 1963; Inhelder and Piaget, 1958; Jenkins and Ward, 1965; Ward and Jenkins, 1965).

Statistically, it is important to distinguish between eliciting an expert's *beliefs about* a population correlation coefficient and eliciting *the value of* the correlation in the expert's beliefs about two variables. In the first case, the correlation coefficient is the variable whose probability distribution we wish to elicit. This task is addressed, for instance, by Gokhale and Press (1982). The second case is the situation that arises in multivariate elicitation, where the correlation (or some other measure of association) is to be elicited as one summary of the expert's joint distribution for two variables. Where the two variables can be considered as single draws from a population, then some of the methods described above may be appropriate. However, many cases of multivariate elicitation do not fit this situation. Consider, for example, eliciting someone's beliefs about the fuel

economy and the acceleration of a new car. This car is not some random draw from any population. Methods 4 to 6 in Clemen et al's study are no longer appropriate.

None of the methods examined by Clemen et al use graphs in any way. It seems likely that graphical methods could perform better, especially as it is very natural to plot a graph to describe the relationship between two variables such as height and weight. This approach would represent association between variables in terms of regression, which is related to correlation. For two variables X and Y, for instance, we might try to elicit the regression function $m(x) = E(Y \mid X = x)$. If the expert accepts the proposition that this function is linear, we might simply elicit $m(x_1)$ and $m(x_2)$ for any $x_1 \neq x_2$. Eliciting more than two points on the function ('over-fitting', see section 4.1) would allow the assumption of linearity to be checked, or a more accurate fitting of a straight line. Here as elsewhere, it may be preferable to elicit medians than means.

A body of psychological research has examined multiple regression. In this research, the x-variables are generally referred to as cues, Y is termed the criterion and the regression coefficients are referred to as cue-weights. Subjects predict the value of the criterion, basing their predictions on the known values of the cues. It has been found, in a wide variety of situations, that a subject's responses can be represented quite well by a linear model that relates the criterion to the cues. The correlations between subjects' responses and the responses predicted by linear models (fitted to the same responses that determined the model) have generally taken values in the 0.70's when the judgemental task is from a "real world" situation, and in the 0.80's and 0.90's for less complex artificial tasks. In some studies, the model derived from one sample of predictions was used to forecast a second sample of predictions. The forecasts produced in this way were only slightly less accurate than those produced by a model actually based on the second sample of predictions (Einhorn, 1971; Slovic and Lichtenstein, 1968; Wiggins and Hoffman, 1968). Experiments also show that, provided cues are monotonically related to the predicted variable, a simple linear combination

of main effects will do a remarkably good job at forecasting a subject's assessments, even if subjects know that interactions exist. One implication is that, when eliciting the dependence of one variable on one or more other variables, it is reasonable to constrain an expert's assessments to fit a linear model and to ignore interactions, unless it becomes clear that some interactions are important. An extensive review of cue-weighting experiments is given in Slovic and Lichtenstein (1971).

A joint distribution involves more than modelling conditional means or medians. Hence, the task of eliciting a joint distribution is more complex than determining an expert's cue-weights. Generally, conditional probabilities are a natural way to augment marginal probabilities when trying to specify a joint probability distribution, and in particular allow conditional dispersion to be elicited and modelled. Conditional medians and other quantiles are extensively exploited, for instance, by Kadane et al (1980), Dickey et al (1986) and Kadane (1996).

An alternative to conditional probabilities is joint probabilities. Having elicited $P(X \leq x)$, for instance, we might elicit the joint probability $P(X \leq x, Y \leq y)$ or, equivalently, the conditional probability $P(Y \leq y \,|\, X \leq x)$. Note, however, that conditional probabilities are usually elicited in the form of $P(Y \leq y \,|\, X = x)$, and conditioning on $X \leq x$ rather than on $X = x$ may be cognitively more complex. On the other hand, we have already noted that experts do not assess joint probabilities accurately, even when the variables are independent. We might also expect joint probabilities like $P(X \leq x, Y \leq y)$ to be subject to a kind of representativeness bias, and so be positively/negatively biased if the association between X and Y is positive/negative, although this does not appear to have been studied.

3. Fitting a Distribution

Once the facilitator has obtained from the expert a number of specific statements, the elicitation task is completed by converting these into a probability distribution. Different levels of complexity are found in the fitting of a probability distribution to the expert's statements. If the elicitation is to obtain a prior distribution which will then be updated in a Bayesian analysis of some additional data, it is usual to fit a probability distribution using standard parametric families of distributions. However, where the elicitation is to formulate uncertainty about inputs to a decision problem or a mathematical model, such as in the risk assessment of a complex engineering project, much more simplistic elicitation and fitting are common.

3.1. Uniform and triangular distributions

The simplest form of elicitation is to ask the expert to specify a range $[a, b]$ in which the parameter is believed to lie. If this is all that is elicited from the expert, then it is common to assume a uniform probability distribution over $[a, b]$. This can be criticised as too simplistic in at least two respects. First, the expert almost certainly would not believe that the unknown quantity in question is as likely to be very close to the limits a and b as to be at a more central point in the interval. Second, unless the range $[a, b]$ represents absolute physical limits to the possible values of the quantity (in which case the first criticism applies even more strongly), it is unreasonable to give zero probability to the event that the quantity lies outside the range.

As a simple response to the first criticism, another common practice is to use a triangular distribution. For this purpose the expert is asked also to specify a mode, say c. Then the assumed

distribution has the density

$$f(x) = \begin{cases} 2\frac{x-a}{(b-a)(c-a)} & \text{if} \quad a \le x \le c \\ 2\frac{b-x}{(b-a)(b-c)} & \text{if} \quad c \le x \le b \end{cases} .$$

The acceptability of uniform and triangular distributions as representations of uncertainty about model inputs in engineering applications is indicated by their featuring strongly in Oberkampf et al (2004), but O'Hagan and Oakley (2004) criticise this practice as a failure to elicit adequately.

Even where substantially more information is elicited from the expert, uniform distributions may be assumed over intervals. Suppose that in addition to the range $[a, b]$ the expert also specifies probabilities p_1, $p_2, \ldots, p_k$ that the uncertain quantity lies in the intervals $[a, c_1]$, $(c_1, c_2], \ldots, (c_{k-1}, b]$. (This may be done by fixing the c_is and asking for the probabilities, or by fixing the p_is and asking for quantiles.) Then the facilitator may simply assign the histogram distribution

$$f(x) = \frac{p_i}{c_i - c_{i-1}} \quad \text{if} \quad c_{i-1} \le x \le c_i \ , \ i = 1, 2, \ldots, k \ ,$$

where $c_0 = a$ and $c_k = b$. Although the expert's beliefs would generally be better represented by a distribution with smooth density function, this histogram form may be adequate, particularly if k is not small. Feedback to the expert may be useful at this point.

3.2. Fitting parametric distributions

More complex elicitation methods usually impose structure on an expert's opinion by assuming that his or her knowledge can be well-represented by some member of a specified family of distributions. If the expert has specified information that fits a convenient parametric distribution, it makes sense to use it. This strategy has costs and benefits similar to those for statistical modeling in general.

Members of the hypothesized family are distinguished by parameters (called *hyperparameters*) and the elicitation task then reduces to choosing appropriate hyperparameter values to capture the main features of the expert's opinion. If the expert's opinion does not correspond approximately to any member of the family, discrepancies are likely to show up in the expert's answers to elicitation questions, much as a sampling model can succumb to traditional diagnostic checks. The family of distributions is typically chosen to be the natural conjugate family (or a tractable extension of that family), which facilitates subsequent analysis if sample data become available, although advances in Bayesian computation through MCMC methods make it viable to use other families. For many sampling models, the conjugate family is reasonably flexible and can represent a variety of opinion through suitable choice of hyperparameters. Further flexibility is available through mixtures of conjugates. Dalal and Hall (1983) and Diaconis and Ylvisaker (1985) demonstrate that such mixtures can arbitrarily accurately represent any actual belief, although we are not aware of any work in which an expert's opinion is elicited in terms of a mixture, other than for variable selection problems (Garthwaite and Dickey, 1992, 1996).

Two elicitation tasks that have attracted substantial attention are quantifying opinion about a Bernoulli process and quantifying opinion about a linear regression model. We first focus on each of these problems in turn before briefly discussing elicitation for other sampling models. The judgemental tasks that are central in much of this work are the assessment of means, medians and quantiles, revision of opinion when sample data becomes available, and specifying relevant aspects of a 'prior sample' whose information content would approximately equate to one's knowledge. As a guiding principle, experts should be asked questions about quantities that are meaningful to them. This suggests that questions should generally concern *observable* quantities rather than unobservable parameters, although questions about proportions and means might also be considered suitable, as psychological research suggests that people can relate to these quantities. However, in

some application areas, particular statistical models are so familiar to experts that their parameters have acquired well understood scientific meaning. It may then be appropriate to ask experts directly about such parameters, as discussed in Kadane (1980) and Winkler (1980). The facilitator should always try to understand what terms make the expert most comfortable for elicitation.

Four basic methods have been used to quantify subjective opinion about p, the unknown parameter of a Bernoulli process. In illustrating assessment questions we shall suppose that p is the proportion of students at the University of Chicago who are male, which is the example used in Winkler (1967), the first paper to address this elicitation problem.

1. One method is to ask the expert to specify her median estimate of p and to give one or more quantiles (usually at least two) of her subjective distribution for p. These may be plotted and a smooth cumulative distribution function drawn through them, giving a nonparametric representation of the expert's opinion. More commonly, it is assumed that the expert's opinion can be well-represented by a beta distribution (the conjugate distribution for Bernoulli sampling), and a beta distribution is selected whose quantiles are similar to those the expert gave. The beta distribution might be selected using a table presented in Winkler (1972, Table 5) that lists several quantiles for a variety of parameter values. We shall refer to this method of elicitation as the *quantile method*; it is also often called the credible interval method.

2. The second method is the *hypothetical future sample (HFS) method*. The expert first estimates the proportion under consideration (e.g. the proportion of students who are male) and then revises her opinion in the light of information from additional (hypothetical) samples. For example, she might be asked questions of the form: "Suppose a random sample of 50 students were taken and 20 of them were male. Now what is the probability that one additional student, chosen at random, is male?" Again it is assumed that the expert's opinion corre-

sponds to a beta distribution; its parameters are uniquely determined by the expert's prior and posterior (given the hypothetical sample data) estimates of the proportion. In general, the expert is confronted with several hypothetical samples in which the number of students and the proportion of these who are male is varied. Each yields a separate estimate of the hyperparameters and some form of averaging is used to reconcile them. A general issue arises here in the case that the resulting hyperparameter estimates are very disparate. This may reflect assessment inaccuracy, i.e. the beta distribution may be correct and the elicitation method may be the best available, but the expert's answers to questions are subject to substantial variability. On the other hand, it also warns that the elicited distribution could be a poor representation of the expert's opinion, either because her opinion does not correspond to a beta distribution, or because the elicitation method asks questions that are hard to answer accurately. See also the discussion of internal consistency in Section 4.1.

3. The third method is the *equivalent prior sample (EPS) method*, in which an expert expresses her knowledge as an equivalent prior sample. Quoting Winkler (1967, p. 779), "You have some knowledge concerning University of Chicago students. Can you determine two numbers r and n such that your knowledge would be roughly equivalent to having observed exactly r males in a random sample of n University of Chicago students, assuming that you had very little knowledge about the sex of University of Chicago students before seeing this sample?" The prior distribution is taken to be a beta distribution with parameters r and $n - r$.

4. The fourth method is the *probability density function (PDF) method*, in which the expert specifies the most likely value of p, say $\widehat{p}$, and then assesses other points of the p.d.f. for p. For example, Winkler (1967) asks the expert to give two values of p (one on each side of $\widehat{p}$) that she considers half as likely as $\widehat{p}$, and to also specify quantiles, which in this context are defined

as points that divide the area under the graph of the p.d.f. in specified proportions. (There is obviously strong similarity between the PDF method and the quantile method, as both make use of quantile assessments.) A nonparametric estimate of the expert's opinion may be obtained by asking her to draw a graph of her p.d.f., taking into account the assessments she has given.

All four of the above methods tend to produce prior distributions that are unrealistically "tight" (i.e. they have variances that are too small). For example, Schaefer and Borcherding (1973) conducted an experiment in which 22 subjects used the EPS and quantile methods to quantify their opinions about various proportions. Before any training, each subject assessed 50% central credible intervals for eighteen different proportions using each method. The proportion of these intervals that contained true values was 15.7% for the EPS method and 22.5% for the quantile method. The task in the HFS method, revising opinion in the light of additional data, is similar to the task in the "bookbag-and-pokerchips" experiment, where conservatism has a marked effect. Insufficient revision of opinion when given the hypothetical data would result in a distribution that is too tight. Winkler (1967) felt that conservatism also influences the EPS method, hypothesizing that subjects equate their knowledge to too large a sample size, through not realizing the value of sample information. The quantile method tends to yield distributions that are again too tight, but slightly less tight than the PDF method and much less tight than the HFS and EPS methods (Winkler, 1967). On this basis the quantile method seems preferable, and it also seems the method of choice when judged by scoring rules. (Scoring rules are discussed in Section 4.3). Some experiments have examined the effect of training and these found that the bias of tightness was reduced for all three methods with particularly marked improvement for the EPS method (Stael von Holstein, 1971; Schaefer and Borcherding, 1973).

Elicitation methods have also been devised that avoid direct questions about p (which is not an

observable quantity) and instead ask questions about sampling distributions, such as "the number of students who would be male in a random sample of twenty University of Chicago students". Chaloner and Duncan (1983) use this form of question in a variant of the PDF method. The expert states the most likely number of males in the sample, x say, and then assesses the relative likelihood of x males rather than $x-1$ males, and of x males rather than $x+1$ males. Assessments are used to estimate the parameters of a beta distribution and the implied endpoints of the shortest 50% predictive interval are then calculated. These endpoints are given as feedback to the expert, who comments on the length of the interval. If she finds it too short or too long, the parameter estimates are revised repeatedly until she is satisfied with its length. The process is repeated for a variety of sample sizes and the resulting parameter estimates are amalgamated in some way. Gavaskar (1988) also uses questions about the sampling distribution in a variant of the HFS method. The expert first specifies the most likely number of males in a random sample of some specified size (as in the method of Chaloner and Duncan) and then revises her assessment after being given hypothetical sample data, as in the HFS method. Assessments are again used to determine the parameters of a beta distribution. This is done for a variety of sample sizes. Interestingly, Gavaskar conducted a computer simulation to assess the sensitivity of parameter estimates to errors in an expert's assessments. He compares his method with that of Chaloner and Duncan (1983) and finds that his method is much less sensitive. However, in a simulation study of this nature it is difficult to choose appropriate error distributions for the different methods and he may have underestimated the magnitude of errors that are induced by hypothetical samples.

Turning to multiple linear regression, we suppose the model is $Y = \mathbf{x}'\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. The usual conjugate prior distribution specifies that σ^2 equals $\omega\delta$ times the reciprocal of a chi-squared random variable with δ degrees of freedom, and that given σ, β has a normal distribution with some mean $\mathbf{b}$ and some variance-covariance matrix $\sigma^2\mathbf{R}$. For this prior distribution, ω, δ, $\mathbf{b}$

and $\mathbf{R}$ are the hyperparameters that need to be assessed via an expert's assessments.

To quantify opinion about these quantities, Zellner (1972) suggests questioning an expert about the regression coefficients. Some experts may be able to think of regression coefficients directly but, as noted earlier, it is usually better to question people about observable quantities, such as Y, rather than ask direct questions about unobservable quantities, such as regression coefficients. To this end, elicitation methods generally ask the expert about observations $Y_1, \ldots, Y_m$ at values $\mathbf{x}_1, \ldots, \mathbf{x}_m$ of $\mathbf{x}$. This has the possible disadvantage though, that uncertainty about Y results both from the expert's uncertainty about the values of the regression coefficients *and* from random variation. To separate these sources of uncertainty, Garthwaite and Dickey (1988) ask questions about $\overline{Y}_i$, the average value of Y if a large number of observations were taken at a single value, $\mathbf{x}_i$, arguing that averages are quantities to which people can relate, and that an expert can give assessments about $\overline{Y}_i$ without the need to consider random error. We briefly outline the elicitation methods that have been proposed by Kadane, Dickey, Winkler, Smith and Peters (1980), Oman (1985), and Ibrahim and Laud (1994), which all question an expert about Y, and the method of Garthwaite and Dickey (1988). In this section we refer to these papers as KDWS&P, Oman, I&L, and G&D, respectively.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$. We refer to $\mathbf{x}_i$ as a design point and $\mathbf{X}$ as a design matrix. To estimate the mean vector, $\mathbf{b}$, at each design point KDWS&P elicit the median of Y, G&D elicit the median of $\overline{Y}$, Oman elicits the mean of Y, and I&L elicit the expert's 'best guess' (or prior prediction) of Y. Denoting the vector of assessments by $\mathbf{y}$, the methods equate $\mathbf{b}$ to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, which would be the least squares estimate if $\mathbf{y}$ were a vector of observations at $\mathbf{X}$. (Oman also offers two less attractive alternative methods of eliciting $\mathbf{b}$: asking the expert to assess $\mathbf{b}$ directly, and asking the expert to specify prior expected covariances.) With each elicitation method, the distribution of Y at any design point is unimodal and symmetric so that, in principle, it should not matter which

point estimate of Y is assessed. In practice though, the distribution of Y may be skew. Then the (unanswered) question arises as to which feature of a skew distribution should correspond to the point of symmetry of a symmetric distribution that is used to represent it.

To elicit ω and δ, G&D and I&L elicit assessments that depend only upon experimental error. G&D ask the expert to suppose that two observations are taken at the same design point. It is pointed out to the expert that the observations will not be identical because of random variation, and the median of their absolute difference is elicited. The expert is then given a hypothetical datum and she then states her updated median of their absolute difference. I&L use assessments of the mean and variance of the precision (σ^{-2}) to determine ω and δ and, in related work (Laud and Ibrahim, 1995), they use assessments of the median and the $95th$ percentile of the distribution of the precision. Oman does not elicit ω and δ, and restricts his posterior analysis to inferences that depend only on a point estimate of σ^2, which he obtains using empirical Bayes methods. KDWS&P elicit δ by asking the expert to assess the median ($y_{.50}$), upper quartile ($y_{.75}$)and 93.75 percentile ($y_{.9375}$) at a design point. ($y_{.9375}$ is elicited through repeated bisections; $y_{.50} \longrightarrow y_{.75} \longrightarrow y_{.875} \longrightarrow y_{.9375}$.) The ratio $(y_{.9375} - y_{.50})/(y_{.75} - y_{.50})$ depends only on δ and hence provides an estimate of it. KDWS&P repeat the assessments at several design points and average the estimates of δ that each set of assessments yields. The method used by KDWS&P to elicit ω is complex and the reader is referred to their paper for details. A central task is to ask the expert to suppose that two independent observations are taken at a specified design point. The median of one of the observations is elicited; then the expert is given a hypothetical value for that observation and assesses the conditional median of the other observation.

All the above methods may be criticized for using assessment tasks that people are not very good at performing. Direct questions about the distribution of a precision (I&L) are surely hard to answer; G&D use conditional assessments that will be biased by conservatism and in addition they

only elicit the minimum number of assessments to determine the hyperparameters; KDWS&P found that extreme values of δ were not uncommon with their method. As δ is a difficult hyperparameter to assess, it is a good idea to elicit more than one estimate of it and to then reconcile the different estimates in some way. If estimates of δ are to be combined arithmetically, empirical evidence favours taking their geometric mean, rather than their average (Al-Awadhi and Garthwaite, 2001).

The most complicated hyperparameter to elicit is the variance-covariance matrix, $\sigma^2\mathbf{R}$; this may contain a large number of elements that each need to be elicited and the matrix itself must be positive-definite. In KDWS&P, a crucial step in assessing this matrix is the elicitation of a variance-covariance matrix for a multivariate t-distribution. We next describe that part of their method in some detail because it can be useful in a variety of elicitation problems. Also, it requires sophisticated mathematics and statistics, so it illustrates that statisticians need to be involved in the development of elicitation methods; in the past some statisticians have suggested that their development should be left to psychologists.

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ has a multivariate t-distribution on δ degrees of freedom, $\mathbf{Y} \sim t_\delta(\mathbf{a}, \mathbf{P})$. Then $\mathbf{P}$ is referred to as the *spread* of $\mathbf{Y}$, $S(\mathbf{Y})$ and, if $\delta > 2$, the variance of $\mathbf{Y}$ equals $[\delta/(\delta-2)]\mathbf{P}$. (The variance does not exist if $\delta < 2$.) To elicit $\mathbf{P}$, KDWS&P proceed as follows.

The expert assesses:

(a) His or her medians of Y_1, ..., Y_m, which we denote by y_1^*, ..., y_m^*.

(b) The upper quartile of Y_1, denoted by $y_{1,.75}$.

(c) Conditional medians. A conditional set of values is built up in stages and, at the ith stage ($i = 1, \ldots, m-1$), it consists of values $y_1^0, \ldots, y_i^0$. Given $Y_1 = y_1^0, \ldots, Y_i = y_i^0$, the expert gives the conditional median of Y_j for $j = i+1, \ldots, m$, the assessment being denoted $y_{j,.50} \,|\, y_1^0, \ldots, y_i^0$.

(d) Conditional quartiles. Given $y_1^0, \ldots, y_i^0$, the expert assesses $y_{i+1,.75} \,|\, y_1^0, \ldots, y_i^0$, the conditional

upper quartile of Y_{i+1}.

In elicitation methods, the most widely used method of estimating variances or spreads from quartile assessments is to divide the assessed interquartile range or semi-interquartile range by the corresponding range of a standard distribution. Let $t[\delta, 0.75]$ denote the semi-interquartile range of a univariate t-distribution on δ degrees of freedom. The variance of Y_i does not exist if $\delta < 2$ so KDWS&P determine spreads,

$$S(Y_1) \;=\; (y_{1,.75} - y_1^*)^2/(t[\delta, .75])^2 \tag{1}$$

and

$$S(Y_{i+1} \,|\, y_1^0, \ldots, y_i^0) \;=\; \frac{(\{y_{i+1,.75}|y_1^0, \ldots, y_i^0\} \;-\; \{y_{i+1,.50}|y_1^0, \ldots, y_i^0\})^2}{(t[\delta + i, .75])^2}, \tag{2}$$

for $i = 1, \ldots, m-1$. The order of the conditional assessments enables conditions to be based on an expert's earlier answers; y_1^0 is set equal to $y_1^* + \frac{1}{2}\{S(Y_1)\}^{1/2}$ and, for $i = 2, \ldots, m-1$, y_i^0 is set equal to $y_{i,.50}|y_1^0, \ldots, y_{i-1}^0 + \frac{1}{2}\{S(Y_i|y_1^0, \ldots, y_{i-1}^0\}^{1/2}$.

An iterative method is used to calculate $\mathbf{P}$. Let $\mathbf{P}_i$ denote the covariance matrix of $Y_1, \ldots, Y_i$ and put $\mathbf{P}_1 = S(Y_1)$. The following equations give $\mathbf{P}_{i+1}$ after $\mathbf{P}_1, \ldots, \mathbf{P}_i$ have been estimated. Put

$$\mathbf{L}_{i+1} \;=\; \{(y_{i+1,.50}|y_1^0) - y_{i+1}^*, \ldots, (y_{i+1,.50}|y_1^0, \ldots, y_i^0) - y_{i+1}^*\}', \tag{3}$$

$$\mathbf{T}_{i+1} \;=\; \begin{pmatrix} y_1^0 - y_1^* & (y_{2,.50}|y_1^0) - y_2^* & \ldots & (y_{i,.50}|y_1^0) - y_i^* \\ y_1^0 - y_1^* & y_2^0 - y_2^* & \ldots & (y_{i,.50}|y_1^0, y_2^0) - y_i^* \\ \vdots & \vdots & \ddots & \vdots \\ y_1^0 - y_1^* & y_2^0 - y_2^* & \ldots & y_i^0 - y_i^* \end{pmatrix}^{-1} \mathbf{L}_{i+1} \tag{4}$$

and

$$S(Y_{i+1}) = \frac{S(Y_{i+1}|y_1^0, ..., y_i^0) \,.\, [1 + i/\delta]}{1 + \delta^{-1}(y_1^0, ..., y_i^0)' \mathbf{P}_i^{-1} (y_1^0, ..., y_i^0)} + \mathbf{T}_{i+1}' \mathbf{P}_i \mathbf{T}_{i+1}. \tag{5}$$

Then put

$$\mathbf{P}_{i+1} \quad = \quad \begin{pmatrix} \mathbf{P}_i & \mathbf{P}_i \mathbf{T}_{i+1} \\ \mathbf{T}_{i+1}' \mathbf{P}_i & S(Y_{i+1}) \end{pmatrix}. \tag{6}$$

The procedure stops when $\mathbf{P}_m = \mathbf{P}$ has been obtained; results in KDWS&P show that $\mathbf{P}$ is certain to be a positive-definite matrix.

A general feature of the method of KDWS&P is that for every hyperparameter it elicits more assessments than necessary, and then uses some form of averaging to obtain hyperparameter estimates. This is obviously sound practice. Also, the largest deviations from the averages are reported to the expert, so that she can judge the extent to which some of her answers may be in error and require changing. The **x**-values that are used in the elicitation method will affect the quality with which expert opinion is captured and Kadane and Wolfson (1998) suggest a procedure for their selection.

Oman does not attempt to relate $\mathbf{R}$ to the expert's opinion. Instead, he chooses design points that cover the region of interest and that give a design matrix $\mathbf{X}$ that is as near orthogonal as possible. He then sets $\mathbf{R}$ equal to $\tau(\mathbf{X}'\mathbf{X})^{-1}$ and estimates τ using empirical Bayes methods. I&L are concerned with the analysis of designed experiments. They let $\mathbf{X}$ be the design matrix for which data will be gathered and, like Oman, assume that $\mathbf{R} = \tau(\mathbf{X}'\mathbf{X})^{-1}$. The hyperparameter τ is then chosen to reflect the weight that should be attached to the expert's opinion, relative to the weight that should be attached to the data from the experiment. Hence, I&L choose $\mathbf{R}$ partly to reflect the expert's knowledge, but their approach is pragmatic and is not a serious attempt to determine the prior variance of β. (Otherwise, their approach implies that prior opinion about the

relationship between Y and the x-variables is dependent upon the experiment to be conducted.)

G&D develop an alternative method of eliciting $\mathbf{P}$, the spread of $\mathbf{Y}$. It is based on a novel assessment task that requires the expert first to select the design point at which she can predict Y most accurately, and then to repeat this task several times with an increasing set of restrictions on the x-values she can choose. The method was developed for the use of industrial chemists and exploits their experience of choosing design points to conduct experiments. It is not as flexible as that of KDWS&P, in that it cannot be used with polynomial regression or with x-variables that are factors. However, it can be extended to elicit prior distributions that are suitable for variable-selection problems (Garthwaite and Dickey, 1992). A feature common to the methods of KDWS&P and G&D is that a structured set of sequential questions is used to ensure that $\mathbf{P}$ is a positive-definite matrix.

A drawback of the method of KDWS&P is that for many of its tasks the expert must update her beliefs on the basis of hypothetical data, so assessments are likely to be biased by conservatism. Al-Awadhi and Garthwaite (1998) suggest a modification whereby the diagonal elements of $\mathbf{P}$ are estimated from unconditional assessments. The modification involves scaling $\mathbf{P}$ in such a way that estimates of correlations are unchanged, so $\mathbf{P}$ is still certain to be a positive-definite matrix while the impact of conservatism should be reduced. Al-Awadhi and Garthwaite also suggest estimating the spreads of univariate distributions from assessments of both the lower and upper quartile (KDWS&P use medians and upper quartiles), so that an estimated spread reflects both halves of the subjective distribution. Eliciting both quartiles also enables marked asymmetry to come to light.

Turning to elicitation methods for other sampling models, the above paper of Al-Awadhi and Garthwaite (1998) gives a method of eliciting a conjugate prior distribution for sampling from a multivariate normal distribution. This work is extended in Garthwaite and Al-Awadhi (2001) so

as to elicit a more flexible non-conjugate prior distribution. The methods follow the approach of KDWS&P to assess spread matrices (with the modifications mentioned above) and follow the approach of G&D to elicit degrees of freedom parameters. The method of KDWS&P is also exploited by Dickey et al. (1986) to develop assessment methods for matrix-t models and and a closely related method is used by Garthwaite and Al-Awadhi (2003) for logistic regression. An elicitation method for logistic regression is also given by Chen et al. (1999), using similar ideas to I&L. Chaloner and Duncan extend their method of quantifying opinion about a Bernoulli process (Chaloner and Duncan, 1983) so as to elicit a Dirichlet distribution that represents prior opinion about a multinomial sampling model (Chaloner and Duncan, 1987). Other models that have attracted attention include the proportional hazards model (Chaloner et al., 1993), Weibull lifetime distributions (Singpurwalla and Song, 1987), AR(1) time series models (Kadane et al., 1996), and ANOVA models (Black and Laskey, 1989). Graphical feedback is an important component in the methods of Chaloner and Duncan (1983, 1986), Chaloner et al. (1993) and Garthwaite and Al-Awadhi (2003) and it seems a potentially powerful means of improving the quality of assessed distributions.

Almost all the above methods represent expert opinion by some form of conjugate distribution. This has limitations, notably when the sampling model is a multivariate normal distribution. In a frequentist analysis for a multivariate normal distribution, the sample variance-covariance matrix is both an estimate of the population variance and, after division by the sample size, it is also the estimated variance of the sample mean. In the standard conjugate distribution, a single variance matrix again fulfils both these roles. An experiment by Al-Awadhi and Garthwaite (2001) demonstrated that this is inappropriate. They examined different forms of assessment task and compared alternative ways of estimating hyperparameters. To quantify opinion about the vector of means, it proved preferable to ask directly about the means rather than individual observations while, to quantify opinion about the variance matrix, it was better to ask about deviations from

the mean. One alternative to the conjugate distribution is to assume that the population mean and variance are independent in the prior distribution, an approach followed in Garthwaite and Al-Awadhi (2001).

Elicitation methods for broader classes of problem have also been proposed. Bedrick et al. (1996) suggest a method for generalised linear models in which the predictive distributions at different design points are elicited and then combined to form a prior distribution. For convenience, Bedrick et al. mainly consider the case where the predictive distributions are independent of each other, so that combining them is straightforward. Properties of their method are discussed and they show it has similarities to data augmentation. Clemen and Reilly (1999) consider the general problem of constructing a joint prior distribution for several hyperparameters. They use a copula to form the joint distribution from an expert's subjective judgements of marginal distributions and correlations. No restrictions are placed on the marginal distributions and dependence between the marginals is modelled by the copula that underlies a multivariate normal distribution. People are not good at assessing correlations and Clemen and Reilly address this problem by discussing various techniques for their assessment. They report a small empirical study that forms part of the basis for their well-informed views; Clemen et al (2000) report a larger study of methods for assessing correlations, as discussed above in Section 1.2.3.

3.3. Nonparametric fitting

Any fitting of a parametric distribution to the expert's stated summaries implies assumptions about the form of the expert's underlying probability distribution. Although the distributional form may be acceptable to the expert, he or she is rarely in a position to question the assumptions critically. This is even more true in the case of multivariate distributions.

Just as there are many statisticians who are uncomfortable with parametric assumptions in

modelling data, it is arguably preferable in elicitation not to make parametric assumptions. (Indeed, in the subjective Bayesian framework, the likelihood is just as much a judgement as the prior distribution, and should in principle also be formally elicited.) A number of nonparametric approaches are possible in elicitation.

One is simply to decline to fit a distribution at all, and to use just the expert's stated summaries and nothing more. It is then necessary to find ways to make use of this limited specification of beliefs. In the context of eliciting a prior distribution for a Bayesian analysis, *Bayes linear* methods have been advocated by Goldstein (1999). The Bayes linear approach is based upon eliciting only first and second order moments (i.e. means, variances and covariances). The underlying theory is a Bayesian analogue of the Gauss-Markov theorem, but the advocates of the Bayes linear approach have developed a substantial body of methodology to facilitate applications. Bearing in mind the difficulties that we have noted earlier in asking experts to assess moments, the Bayes linear approach places higher demands on the statistical understanding of the expert, or else requires a substantial training input.

Berger and O'Hagan (1988) effectively allow the expert's prior distribution for a single unknown parameter to be any unimodal distribution having specified quantiles. They compute the range of posterior inferences over this range of prior distributions, for given data. This is a fully nonparametric approach, although it may be difficult to implement in more complex situations. It also allows distributions that the expert would easily be able to reject as inaccurate representations of his or her knowledge.

A more recent approach is that of Oakley and O'Hagan (2002), which adopts the framework of modern Bayesian nonparametrics. The expert's beliefs about the random variable X are supposed to be represented by a probability density function $f(x)$. To the facilitator f is an unknown function. The facilitator has a prior distribution for f, and this is updated to a posterior distribution

after 'observing' the 'data' comprising the expert's summary values. It is worth looking at this formulation in a little more detail.

Since f is a function, the facilitator's prior is a distribution over the space of possible values of the whole function. Formally, Oakley and O'Hagan (hereafter, O&O) suppose that f has a Gaussian process prior distribution. Its prior expectation is a member $g(x;\theta)$ of some parametric family indexed by θ. The prior variance of $f(x)$ is $g(x;\theta)^2\sigma^2$ and the correlation between $f(x)$ and $f(x')$ is a decreasing function of the distance between x and x'. The hyperparameters θ and σ^2 are given weak prior distributions. The method is nonparametric because f is not assumed to be exactly a member of any parametric family, and indeed it is allowed to take any form at all. Nevertheless, the facilitator expects the expert's true density function f to be *close*, in some sense, to some member of the parametric family defined by g. The closeness is governed by the hyperparameter σ^2, which O&O learn about from the expert's summary 'data'. The correlation in the Gaussian process ensures that departures of $f(x)$ from $g(x;\theta)$ are *smooth*.

Although in this approach the facilitator does not constrain the expert's distribution to fit the parametric family, it is clear that the facilitator is supplying some information through the Gaussian process prior distribution. There is the belief that f should be 'close' to g, and that it should be a smooth density function. Technically, from the assumption of a Gaussian process the facilitator has a normal distribution for each $f(x)$. Strictly, this cannot be completely realistic because $f(x)$ must be non-negative, but the normality is important for tractability of the O&O approach, and they argue that it is relatively innocuous.

A significant benefit of the O&O approach is that it yields a complete posterior distribution for f. The posterior mean can be regarded as an estimate of f based on the expert's summaries, and hence as the elicited probability distribution. But O&O then have a posterior distribution that quantifies the possible inaccuracy of the elicitation (as called for by Dickey, 1980). They give the

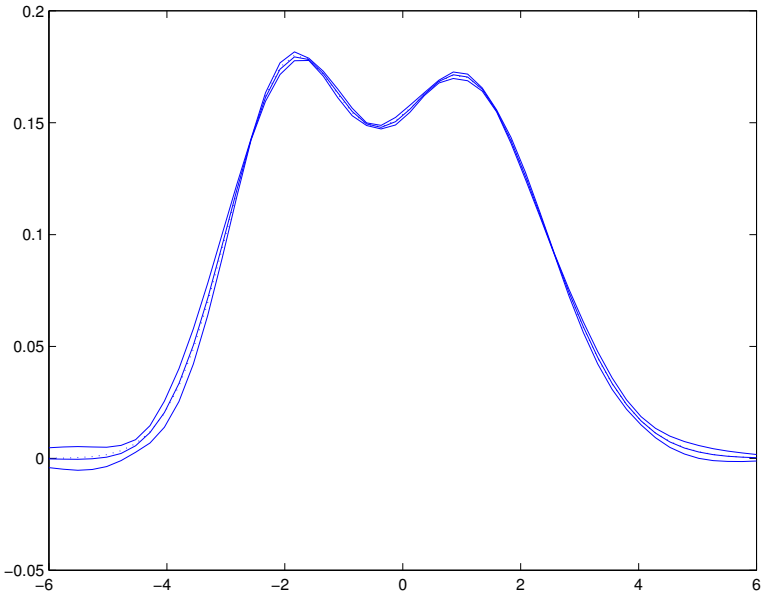

Figure 2: Example from Oakley and O'Hagan (2002). Median and pointwise 95% intervals for the expert's density function (solid lines), and the true density function (dotted line)

following synthetic example in which the expert's true probability distribution is a bimodal mixture of two normal distributions. The parametric family g is the normal family. Figure 2 shows the posterior median and 95% pointwise credible intervals for f based on seven elicited quantiles from the expert. We see that, despite believing initially that the expert's distribution would be similar to a normal distribution, the facilitator's posterior distribution accurately reproduces the expert's true bimodal f.

4. Testing Adequacy of Elicitation

In view of the many practical difficulties of elicitation, how can one know whether the elicited distribution is, in any sense, an *adequate* representation of the expert's knowledge? Before addressing this, we should consider whether there is, in some sense, a 'true' representation. Does the expert have a 'true' personal probability distribution for the uncertain quantities?

Winkler (1967, p.778) writes,

"The assessor has no built-in prior distribution which is there for the taking. That is, there is no 'true' prior distribution. Rather, the assessor has certain prior knowledge which is not easy to express quantitatively without careful thought. An elicitation technique used by the statistician does not elicit a 'true' prior distribution, but in a sense helps to draw out an assessment of a prior distribution from the prior knowledge. Different techniques may produce different distributions because the method of questioning may have some effect on the way the problem is viewed."

On the other hand, O'Hagan (1988) explicitly defines 'true' probabilities as those that would result if the expert were capable of perfectly accurate assessments of her own beliefs. O'Hagan regards different 'stated' probabilities, that might result from different elicitation methods, as more or less inaccurate attempts to specify the expert's underlying 'true' probabilities. In contrast, Winkler's position seems to be that the results of different elicitations are all assessments of slightly different probabilities. A possible reconciliation is that a 'true' distribution would be the result of a method that leads the expert to view the problem from as complete and unbiased a perspective as possible through appropriate use of cognitive tools.

In this section, we first consider how to test the internal consistency of the expert's statements, together with any assumptions made by the facilitator. We then discuss assessing the adequacy of the elicitation, in terms of whether the acknowledged inaccuracies in the elicitation process matter.

4.1. Internal consistency

A system of probability statements is *coherent* if the probabilities are all consistent with the laws of probability. If, for instance, an expert states $P(E) = 0.4$, $P(F) = 0.3$ and $P(E \text{ or } F) = 0.6$, when E and F are mutually exclusive events, then these probabilities are non-coherent. One way to check the quality of an expert's statements is for the facilitator to ask for sets of probability

assessments that allow tests of coherence. We must expect that the expert's elicited statements will fail coherence tests. This is almost inevitable in view of the imprecision with which the expert can make these judgements. The question then arises of how we should reconcile the internal inconsistency of the elicited statements.

In the case of an incoherent set of individual probabilities, as in the example of mutually exclusive events, the simplest answer is to confront the expert with the inconsistency and to invite her to revise one or more of the stated values. In general we should expect this revision to improve the expert's assessments.

A careful examination of reconciling incoherent probability assessments is given by Lindley, Tversky and Brown (1979, hereafter LT&B). In their approach, the reconciliation is done by the person that we have called the facilitator. The facilitator takes a view of the accuracy with which the expert will have been able to assess the stated probabilities. Thus, in the example of mutually exclusive events, the facilitator needs to formulate a joint probability distribution $p(\mathbf{e}, \mathbf{s})$ for the expert's underlying true probabilities $\mathbf{e} = (P(E), P(F))$ and for the expert's stated probabilities $\mathbf{s}$ for the three events E, F and 'E or F'. The facilitator's beliefs about the expert's true probabilities would then be expressed by $p(\mathbf{e} \,|\, \mathbf{s} = (0.4, 0.3, 0.6))$. In practice, LT&B envisage the facilitator formulating the joint distribution via a prior distribution $p(\mathbf{e})$ for $\mathbf{e}$ and a 'likelihood' $p(\mathbf{s} \,|\, \mathbf{e})$ for the expert's assessment errors, so that the facilitator's posterior distribution $p(\mathbf{e} \,|\, \mathbf{s} = (0.4, 0.3, 0.6))$ is derived by Bayes' theorem. LT&B refer to this solution as the 'internal approach', contrasting it with an 'external approach' that we consider in Section 5.1.

As in the simpler method of asking the expert to revise her own probabilities, this reconciliation can lead to more accurate assessments. However, the improvement is now formally expressed by the familiar Bayesian property that the facilitator's posterior distribution of $\mathbf{e}$ will generally be more concentrated than the prior distribution.

In the context of eliciting a parametric probability distribution sometimes only as many summaries are elicited as are required to identify unique values of the required hyperparameters. It is then usually the case that any set of elicited summaries will be consistent with the fitted distribution, and hence with each other. It is therefore not possible to find any non-coherence. However, if the facilitator asks for at least one more summary from the expert, then it becomes possible to test for coherence. This has been called *over-fitting*.

Note, however, that inconsistent statements from the expert may simply indicate that the expert's distribution cannot be adequately represented by a member of the chosen family. For instance, if the expert's beliefs are sought concerning a proportion π, the facilitator might choose to work with the assumption of a beta distribution. Then if the expert specifies that the median is 0.4, the lower quartile is 0.3 and the upper quartile is 0.5, then no beta distribution will fit these specifications. The $Be(4.733, 7.1)$ distribution fits the median and lower quartile, but has an upper quartile of 0.494, while the $Be(4.16, 6.24)$ distribution fits the median and upper quartile, but has a lower quartile of 0.293. However, it is unreasonable to expect the expert to specify these quantiles to such accuracy, and the assessments in practice would not be seen as challenging the assumption of a beta distribution. Instead a compromise such as $Be(4.4, 6.6)$, with lower quartile 0.296 and upper quartile 0.497, clearly fits the elicited values very well.

If on the other hand the expert had specified a lower quartile of 0.2, then the beta family assumption is called into question. The expert's distribution appears to be negatively skewed but with a median below 0.5, two properties which are together inconsistent with a beta distribution. In general, given a set of expert statements that is larger than is needed to identify unique hyperparameters, we can choose an elicited distribution that, in some sense, fits the elicited statements as closely as possible; an example is given in O'Hagan (1998). The quality of fit can be seen as indicating the accuracy of the elicitation, while a sufficiently poor fit casts doubt on the assumed

family of distributions.

In the case of parametric elicitation, then, over-fitting has the potential either to refine the specification of hyperparameters or to refine the assumed distributional family. Notice, however, that to decide between these two options the facilitator needs to have some idea of the accuracy of the expert's judgements. Given that this kind of judgement by the facilitator is required, it may be that an extension of the approach of LT&B might be developed for this case, but we are not aware of any published work in this direction.

The O&O nonparametric elicitation method of Section 3.3 also adopts essentially the approach of LT&B, in the sense that the facilitator derives a posterior distribution for the expert's underlying density function, using the elicited statements as data. Note, however, that O&O treat the expert's summaries as error-free, and so they do not consider an analogue of the likelihood $p(\mathbf{s} \mid \mathbf{e})$.

An idea similar to over-fitting is feedback. For instance, in a parametric elicitation when we have elicited enough summaries to fit a unique member of the chosen family, instead of eliciting one or more further summaries the facilitator informs the expert of the values of those summaries that are implied by the expert's statements so far (and the assumed distributional form). In general, feedback entails displaying the implications of other statements and inviting the expert to confirm or deny that these are reasonable expressions of her beliefs. Whereas over-fitting will almost invariably expose inconsistencies in the expert's statements, feedback often simply results in the expert confirming the implied values. As such, over-fitting is generally preferred, but feedback can be very useful to show complex implications, such as displaying the fitted density function graphically.

It is also worth noting that the expert will often make qualitative statements during the elicitation that can be checked informally against quantitative summaries. For instance, the expert may appear uncertain and have difficulty specifying a credible interval, and yet may actually give

a narrower interval than in another task where they informally indicated more certainty. The facilitator should be alert for any opportunity to assist the expert by checking the consistency of their opinions, whether expressed or implied.

4.2. Fitness for purpose

Although over-fitting and coherence checking have the potential to improve the elicitation process, appreciable imprecision will inevitably remain in the elicited summaries and in the fitted distribution. Whether this imprecision matters depends on the purpose for which the elicitation is performed.

Recognition that an elicited prior distribution does not necessarily reflect the expert's knowledge accurately has led to quite widespread use of sensitivity analysis in Bayesian statistics (O'Hagan and Forster, 2004, chapter 8). This may involve varying the hyperparameters of a parametric fit or more sophisticated variation of all aspects of the distribution. The general thrust of the *robust Bayesian* movement was to allow the true prior distribution to lie in a nonparametric class of distributions containing the elicited distribution. Then this approach proceeded to derive bounds on posterior inferences as the prior varied across the class of possible priors. Berger (1994) reviews this body of research. A more common use of sensitivity analysis in practical Bayesian analyses has been just to explore in an *ad hoc* way a small number of alternative prior distributions.

Whether the elicitation is to obtain a prior distribution for some Bayesian analysis, to obtain expert judgements for inputs of some decision model or for some other purpose, sensitivity analysis has the same objective. It is to determine whether, when the elicited distribution is varied to other distributions that might also be consistent with the expert's knowledge, the results derived from that distribution change appreciably. If not, the elicitation has adequately represented the expert's knowledge.

How can we determine whether the result changes 'appreciably' as the elicited distribution changes? There is a clear answer to this question when the result is a decision that is to be made optimally with respect to some utility function. Then it is not the change in the decision that matters but the change in expected utility. We consider the difference between, on the one hand, the expected utility that is obtained by the optimal decision with respect to the elicited prior and, on the other hand, the maximum expected utility that can be obtained by the optimal decision with respect to any other distribution in the class. This difference represents the potential gain in expected utility that might be obtained by more careful elicitation. See Kadane and Chuang (1978) and Chuang (1984).

The problem with any sensitivity analysis is to specify the class of distributions. If we allow the distribution to vary more from the elicited distribution, then we can expect greater discrepancies in the results. The classes of priors used in robust Bayesian analysis are arbitrary and not based on analysis of the elicitation process. The 'internal approach' of LT&B and the method of O&O both yield the facilitator's posterior distribution for the expert's underlying probabilities or density function. They therefore provide formal expression of the uncertainty around the particular elicited distribution, which can in principle form the basis for subsequent sensitivity analysis. Their formal structures are more complex to apply, but otherwise there seems no alternative to the kind of informal, *ad hoc* sensitivity analysis most commonly employed.

4.3. Scoring Rules

In empirical work, probability distributions may be elicited for uncertain quantities whose actual values are known to the experimenters. In other circumstances, such as weather forecasting, predictive distributions may be assessed for quantities whose values become known subsequently. In both cases it can be useful to compare assessed probability distributions with the observed data to

provide an objective measure of its accuracy. This is the purpose of a scoring rule.

Formally, a scoring rule is a formula for awarding a score to the expert, which can be thought of as a reward. It is a function both of the expert's elicited probability distribution for the uncertain quantity and of that quantity's true value.

One common application of scoring rules is to compare alternative elicitation methods or different variants of an elicitation method. In empirical research, one elicitation method is often judged to be better than another if it gets better scores. Note, however, that better scores result both from the expert assessing her beliefs more accurately and from the expert having more (or more accurate) knowledge.

Other purposes of a scoring rule are to provide an incentive for experts to record their opinions well, and to help train experts to quantify their opinions accurately. To this end, it is important that a scoring rule should encourage experts to record their true beliefs. More precisely, *"The scoring rule is constructed according to the basic idea that the resulting device should oblige each participant to express his true feelings, because any departure from his own personal probability results in a diminution of his own average score as he sees it."* (de Finetti, 1962, p. 359). A scoring rule with this property is termed *proper*. Various proper scoring rules have been proposed and several, including those most commonly used, are described in Matheson and Winkler (1976). See also O'Hagan and Forster (2004, sections 2.54 to 2.58).

4.4. Calibration

There is a large and somewhat murky literature on the subject of calibration. At its simplest, the idea is that perhaps a person's elicited probabilities show a particular flaw, in that, of the events the person says has probability p of happening, some function $g(p)$ of them actually occur. Then the thought is that when the person announces p as their probability of some event, knowing better,

the user of this information has $g(p)$ as their probability (Lichtenstein et al., 1982). Such a program has the following flaw. Suppose the person being elicited is faced with a coin that person believes to be fair, and hence announces $p = \frac{1}{2}$ as the elicited probability of 'tails'. What values can $g(\frac{1}{2})$ take? Since the $g(.)$ are supposed to be probabilities, and 'tails' and 'heads' are mutually exclusive and exhaustive, we must have $g(\frac{1}{2}) + g(\frac{1}{2}) = 1$, i.e. $g(\frac{1}{2}) = \frac{1}{2}$. Now suppose there are three events equally likely, in the view of the person being elicited. The same argument shows $g(\frac{1}{3}) = \frac{1}{3}$ and $g(\frac{2}{3}) = \frac{2}{3}$. Indeed this argument demonstrates $g(r) = r$ for every rational number r. An assumption of continuity or measurability of g then suffices to show $g(x) = x$ for all real numbers, $0 < x < 1$. Hence recalibration on this basis contradicts the coherence of either the pre- or post-transformation probabilities. Note that this argument does not apply to functions g that involve more than the elicited probabilities. For example, it is not a contradiction to coherence to think that a person may be over-confident in the sense that the probability $(\frac{1}{2})$ assigned to the interquartile range is too high, and hence the probability assigned to the tails is too low. Similarly, it might be noticed that a weather forecaster systematically over-predicts rain, and hence under-predicts the event of no rain.

In practice, calibration is relevant where, as in the case of weather forecasters, experts regularly make similar probability statements so that it is possible to check calibration and feedback is immediate, relevant and frequent. Even without a formal calibration check, receiving regular feedback would tend to ensure that their forecasts are reasonably well calibrated.

5. Multiple Experts

5.1. Synthesising separate elicitations

Where important decisions or inferences are to be made, it is common to wish to draw upon the expertise of several experts. A number of approaches have been proposed as to how to elicit, and how to synthesise, the different experts' knowledge. Formal methods of combining probability distributions are reviewed by Genest and Zidek (1986) who give a very useful annotated bibliography, and French (1985) among others. We first consider the situation where the experts do not interact. Separate probability distributions are elicited from the experts, in separate elicitation sessions. It is then natural to ask how we can synthesise these different distributions into a single distribution.

Two of the most popular methods fall into the category known as *opinion pools*. The linear opinion pool is a convex combination (a weighted average) of the individual probability distributions comprising it, and the logarithmic opinion pool is a normalized weighted geometric mean (equivalent to applying a linear pool to the logarithms of the individual probability densities and then normalizing the result). An important property that an opinion pool might be expected to have is that it be externally Bayesian (Madansky 1978), meaning that, when there is an agreed likelihood function, the opinion pool of the posterior distributions should coincide with the posterior distribution obtained from the opinion pool of the prior distributions. Except in trivial cases, the linear opinion pool fails to have this property, while the logarithmic pool does have it, when the weights sum to one. However, a second property that we might require is invariance to event combination. Suppose for instance that we elicit the experts' probabilities for two mutually exclusive events A and B. Letting C be the event 'A or B', each expert's probabilities (assuming they are coherent) satisfy $P(C) = P(A) + P(B)$. Combination invariance would then require that the same property should hold for the pooled probabilities of A, B and C. McConway (1981) shows that

only the linear opinion pool satisfies a general marginalization criterion of this type. It is therefore not possible to find a mechanistic opinion pooling method that both is externally Bayesian and satisfies the marginalization criterion.

Note that the logarithmic opinion pool also suffers from the fact that a single expert's opinion that a certain set has probability zero implies that the pool must also assign zero probability to that set. See Genest and Zidek (1986) for a wide-ranging discussion of these issues.

Both linear and logarithmic pools allow different weights to be assigned to the experts, which can be used to give more weight to experts whose probability distributions are believed to be more accurate. Cooke (1991) describes a method of choosing weights based on the experts' performance in assessing distributions for *seed variables*, which are quantities whose true value is known to the facilitator but not to the experts. Evidence that this produces better elicitation than simple equal weighting of the experts is presented in Cooke and Goossens (2000).

Mechanistic pooling methods can lead to a form of double counting of expertise if the knowledge of some of the experts overlaps substantially. Then it is inappropriate to weight them all equally with other experts, but the method of seed variables will also tend to overweight such a group.

Another criticism of all these pooling methods is that it is not clear whose opinion (if anyone's) the resulting probability distribution represents. A quite different approach to putting multiple experts' opinions together is to imagine each opinion as data input to a single 'supra Bayesian', who uses these opinions to update his or her views. This is the 'external approach' of Lindley, Tversky and Brown (1979). It was proposed earlier by Morris (1974) and is further developed by Lindley (1985), French (1985), Genest and Schervish (1985); see also the discussion of Genest and Zidek (1986). This approach is completely Bayesian, but requires a very substantial elicitation of the supra Bayesian's opinions about the expert opinions to be pooled.

5.2. Group elicitation methods

We now consider approaches where the experts interact as a group. One simple and practical group elicitation approach is to bring the experts together to discuss the uncertain quantity or quantities about which their beliefs are to be elicited, and through this sharing of their expertise to seek a consensus view. In effect, this treats the group as a single individual. Phillips (1999) presents a formal justification of this *behavioural aggregation* approach. There are, however, new psychological issues that arise in the interaction between the members of the group. This kind of group elicitation requires a knowledgeable and experienced facilitator who needs to be aware of the possibilities of strong personalities in the group having too much weight in the discussion, or of judgements based on overlapping experience being overweighted through being repeated in the discussions. It may also be that the pressure to reach consensus leads to the experts suppressing dissenting views, or alternatively it may not be possible to reach consensus.

The Delphi method is a formal technique for managing the group interaction. The method proceeds by first eliciting the various experts' opinions separately, then feeding each expert's views to all the other experts along with some explanation of that expert's reasoning. The experts are then invited to revise their own elicitations. The method then operates iteratively, feeding back the revised elicitations to all the experts, with explanation of the reasons for revisions, and so on. Although some of the complications of group interaction are removed, the method is likely to produce a less efficient sharing of knowledge than the behavioural aggregation approach. It is also still necessary for the facilitator to manage the interaction, since one expert's reasons may have undue influence if very forcefully expressed. Pill (1971) reviews the Delphi technique. In addition, there is a truly vast literature on its use in political science and government.

A variant of Delphi is discussed by DeGroot (1974), who proposes that each expert revises

his/her opinion by applying a linear opinion pool to all the experts' distributions, with weights that reflect the importance that each pool member puts on the opinions of each of the other participants. The system of revisions then forms a Markov Chain whose transition matrix is the matrix of weights, and DeGroot obtains conditions for them to converge to a consensus. See also Lehrer (1976).

In contrast to these practical group elicitation methods, we can also take a more axiomatic approach and try to identify what would be rational ways for experts to seek a combined expression of beliefs. Bayesian theory is profoundly a theory of rational individual decision-making. Basically, it imposes a minimal condition on an individual's statements of what bets would be acceptable (namely the avoidance of Dutch Book), and accepts all responses that meet that condition. How can this theory be extended to groups?

In order to make progress on this question, one needs to know how the group decision-making structure works, and how it relates to the views of the individuals in the group. Before starting, it is necessary to eliminate one obvious special case, that of a dictatorship. In a dictatorship, the choices of only one individual matter for the group's choices, namely those of the dictator. Hence if the dictator behaves individually in accordance with the Bayesian axioms, the group will as well. By eliminating this case, we insist that more than one person's views matter in how the group makes decisions.

Suppose instead the group makes decisions by majority vote, and suppose each member of the group has a transitive ranking of the alternatives. Surely in this case, something reasonable should be true of the decisions of such a group. Well, perhaps not. Suppose the group consists of three people (the minimum for interesting majority votes), and are expressing preferences among three alternatives, A, B, and C. Suppose voter 1 ranks them in that order, i.e. prefers A to B to C. These preferences will be denoted $A > B > C$. Suppose voter 2's preferences are $B > C > A$, and

voter 3's are $C > A > B$. In a choice between A and B, voters 1 and 3 prefer A to B. Between B and C, voters 1 and 2 prefer B to C. Finally, between A and C, voters 2 and 3 prefer C to A. Hence the majority preferences can be summarized as $A > B > C > A$. No group utility function can summarize such choices because they are not transitive. This simple example has been generalized to every non-dictatorial group decision-making process in a celebrated theorem of Arrow (1951, 1963). A huge literature has grown up around this result, mainly under the heading of political economy.

A second approach to the problem of Bayesian group decision-making asks for preferences between risky outcomes for two or more Bayesians. However they make their decisions, these Bayesians seek to compromise. The only condition imposed on their compromises is that they obey the Pareto Principle: if each member of the group prefers A to B, the compromise cannot prefer B. Suppose there are two Bayesians and three alternatives. Two trivial cases must be dealt with first. If the Bayesians agree in their probabilities, every non-trivial convex combination of their utilities, together with their agreed probability, will provide a Bayesian compromise satisfying the Pareto Principle. Similarly, if their utilities coincide, every non-trivial convex combination of their probabilities will similarly suffice. The result of Seidenfeld, Schervish and Kadane(1989) is that these are the only cases in which a Bayesian compromise can be found. When there are more than two Bayesians involved, the Pareto Principle is less binding, so in certain cases a Bayesian compromise can be found (see Goodman, 1988).

These results cast serious doubt on what might be meant by the probability and utility function of a group seeking to be Bayesian. In what sense can the probability and utility function of the group be representative of the decisions the group might make?

6. Discussion

From the 1960s to at least the early 1980s, research in elicitation was substantial, and was characterised by some close collaborations between statisticians and psychologists. More recently, there seems to have been much less research in both the statistics and psychology communities, and collaboration between them has lapsed. There are, however, signs that this is changing. The growing sophistication of Bayesian computational methods has led to a dramatic increase in the breadth and complexity of Bayesian applications. Bayesians are beginning to show more interest in elicitation and, being freed from the computational constraint to use tractable, conjugate priors, there is a need to develop processes capable of eliciting complex, non-standard distributions. A recent review of case studies and software is Kadane and Wolfson (1998). This in turn is likely to lead to renewed collaboration with psychologists.

Despite the existence of a broad and diverse literature in elicitation, which has provided many valuable procedures and insights, there remains very considerable scope for further research. Some important topics in the authors' opinion are the following.

- Multivariate elicitation. Where it is necessary to elicit a joint probability distribution for two or more quantities, there has been relatively little investigation by psychologists (particularly when the quantities cannot be regarded as instances of a larger population). In this context, also, the available multivariate parametric families typically impose unrealistic constraints on experts' beliefs.

- Nonparametric elicitation. We find the use of uniform, triangular or histogram distributions unrealistic, but fitting parametric distributions also imposes constraints that may be unrealistic. There has been little work on nonparametric fitting, and although a nonparametric fit might represent an expert's beliefs more accurately it is not clear whether this will actually

matter in practice (see the discussion in Section 4.2).

- Graphical tools. We believe there is substantial, as yet relatively unexplored, potential in graphical methods.

An aim of much statistical research is to wring as much from data as we possibly can, but using expert opinion better (or using it at all) could add more information than slight improvement in efficiency through better techniques of data analysis. Too often, *ad hoc* methods must be employed when an expert's opinion is to be quantified; ideally, there should be a range of tried and tested elicitation methods in a statistician's toolbox.

Bibliography

Al-Awadhi, S. A. and Garthwaite, P. H. (1998). An elicitation method for multivariate normal distributions. *Communications in Statistics: Theory and Methods* **27**, 1123–1142.

Al-Awadhi, S.A. and Garthwaite, P. H. (2001). Prior distribution assessment for a multivariate normal distribution: an experimental study. *Journal of Applied Statistics* **28**, 5–23.

Alpert, M. and Raiffa, H. (1969). A progress report on the training of probability assessors. *Unpublished manuscript*, Harvard University. Reprinted in *Judgement and Uncertainty: Heuristics and Biases*, D. Kahneman et al. (eds.), 306–334. 1982. Cambridge: Cambridge University Press.

Arrow, K. (1951,1963). *Social Choice and Individual Values.* J. Wiley and Sons, New York.

Barclay, S. and Peterson, C. R. (1973). Two methods for assessing probability distributions. *Technical Report 73-1.* Decisions and Designs, McLean.

Barclay, S. and Randall, L. S. (1975). Interactive Decision Analysis Aids for Intelligence Analysis.

Technical Report Dt/TR 75-4, Decisions and Designs, Inc.: McLean, Va.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance* **9**, 396–406.

Beach, L. R. and Scopp, T. S. (1967). *Intuitive statistical inferences about variances*. Seattle: L. R. Beach, unpublished mimeo.

Beach, L. R. and Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science* **5**, 161-162.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996). A new perspective on priors for generalised linear models. *Journal of the American Statistical Association* **91**, 1450–1460.

Berger, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test* **3**, 5–124.

Berger, J. O. and O'Hagan, A. (1988). Ranges of posterior probabilities for unimodal priors with specified quantiles. *Bayesian Statistics 3*, J. M. Bernardo et al., eds., 45–66. Oxford University Press.

Beyth-Marom, R. (1982). How Probable is Probable? A Numerical translation of verbal probability expressions. *Journal of Forecasting* **1**, 257–269.

Black, P. and Laskey, K. (1989). Models for elicitation in Bayesian ANOVA: implementation and application. *ASA Proceedings of Statistical Computing Section*, pp. 247–252.

Chaloner, K. (1996). The elicitation of prior distributions. *Case Studies in Bayesian Biostatistics*, D. Berry and D. Stangl (eds.), 141–156. New York: Dekker.

Chaloner, K. M., Church, T., Matts, J. P. and Louis, T. A. (1993). Graphical elicitation of a prior distribution for an AIDS clinical trial. *Statistician* **42**, 341–353.

Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a beta distribution: PM elicitation.

Statistician **32**, 174–180.

Chaloner, K. M. and Duncan, G. T. (1987). Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation.*Communications in Statistics: Theory and Methods* **16**, 511–523.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society, Series B* **61**, 223–242.

Chuang, D. T. (1984). Further theory of stable decisions. In *Robustness of Bayesian Analyses*, J. B. Kadane (ed.),165–228. Amsterdam; North Holland.

Clemen, R. T., Fischer, G. W. and Winkler, R. L. (2000). Assessing dependence: some experimental results. *Management Science* **46**, 1100–1115.

Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science* **45**, 208–223.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford: Oxford University Press.

Cooke, R. M. and Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry* **90**, 303–309.

Dalal, S. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society, Series B* **45**, 278–286.

de Finetti, B. (1962). Does it make sense to speak of 'good probability appraisers'? In *The Scientist Speculates – An Anthology of Partly-Baked Ideas,* I. J. Good (ed.), 356–364. London: Heineman.

DeGroot, M. H. (1974). Reaching a Consensus, *Journal of the American Statistical Association* **69**, 118–121.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. In *Bayesian Statistics 2*, J. M. Bernardo et al (eds), 133–156. North-Holland: Amsterdam.

Dickey, J. (1980). Beliefs about beliefs, a theory of stochastic assessments of subjective probabilities. In *Bayesian Statistics*, Bernardo, J., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.),.471-487. University Press, Valencia, Spain.

Dickey, J. M., Dawid, A. P. and Kadane, J. B. (1986). Subjective probability assessment methods for multivariate-t and matrix-t models. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P. Goel and A. Zellner (eds.), 177–195.

Edwards, W. and Phillips, L. D. (1964). Man as transducer for probabilities in Bayesian command and control systems. *Human Judgements and Optimality*, G. L. Bryan and M. W. Shelley, eds. New York: Wiley.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behaviour and Human Performance* **6**, 1–27.

Erlick, D. E. (1964). Absolute judgement of discrete quantities randomly distributed over time. *Journal of Experimental Psychology* **67**, 475–482.

Fischhoff, B. and Beyth, R. (1975). "I knew it would happen"-Remembered probabilities of once-future things. *Organizational Behavior and Human Performance* **13**, 1–16.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1978). Fault trees: sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance* **4**, 330–334.

French, S. (1985). Group consensus probability distributions: a critical survey. In *Bayesian Statistics 2*, J. M. Bernardo et al. (eds.), 83–210. Amsterdam: North Holland and Company.

Garthwaite, P. H. (1983). Assessment of prior distributions fir normal linear models. PhD Thesis. University College of Wales at Aberystwyth.

Garthwaite, P. H. (1989). Fractile assessments for a linear regression model: an experimental study. *Organizational Behavior and Human Performance* **43**, 188–206.

Garthwaite, P. H. and Al-Awadhi, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society, Series B* **63**, 95–110.

Garthwaite, P. H. and Al-Awadhi, S. A. (2003). Quantifying opinion about a logistic regression using interactive graphics. Submitted for publication.

Garthwaite, P. H. and Dickey, J. M. (1988). Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society, Series B* **50**, 462–474.

Garthwaite, P. H. and Dickey, J. M. (1991). An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making* **4**, 17–31.

Garthwaite, P. H. and Dickey, J. M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *Annals of Statistics* **20**, 1697–1719.

Garthwaite, P. H. and Dickey, J. M. (1996). Quantifying and using expert opinion for variable-selection problems in regression (with discussion). *Chemometrics and Intelligent Laboratory Systems* **35**, 1–43.

Garthwaite, P. H. and O'Hagan, A. (2000). Quantifying expert opinion in the UK water industry: an experimental study. *The Statistician* **49**, 455–477.

Gavaskar, U. (1988). A comparison of two elicitation methods for a prior distribution for a binomial

parameter. *Management Science* **34**, 784–790.

Genest, C. and Schervish, M. J. (1985). Modelling expert judgments for Bayesian updating. *Annals of Statistics* **13**, 1198–1212.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography (with discussion). *Statistical Science* **1**, 114-148.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychological Review* **103**, 592–596.

Gokhale, D. V. and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Journal of the Royal Statistical Society A* **145**, 237–249.

Goldstein, M. (1999). Bayes linear analysis. In *Encyclopedia of Statistical Sciences* (update volume 3), S. Kotz et al., eds., 29–34. J. Wiley; New York.

Goodman, J. (1988). Existence of Compromises in Simple Group Decisions, PhD. dissertation, Department of Statistics, Carnegie Mellon University.

Hammerton, M. (1975). A case of radical probability estimation. *Journal of Experimental Psychology* **101**, 242–254.

Hampton, J. M., Moore, P. G. and Thomas, H. (1973). Subjective probability and its measurement. *Journal of the Royal Statistical Society A* **136**, 21–42.

Hofstatter, P. R. (1939). Uber die Schatzung von Gruppeneigenschaften. *Zeitschrift fur Psychologie* **145**, 1–44.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association* **70**, 271–294.

Hogarth, R. M. (1987). *Judgement and Choice* (2nd ed.). Chichester: John Wiley.

Howard, R. A. and Matheson, J. (1984). Influence diagrams. In *Readings on the Principles and Applications of Decision Analysis*, R. A. Howard and J. Matheson (eds.). Strategic Decision Group, Palo Alto, California.

Huber, G. P. (1974). Methods for quantifying subjective probabilities and multi-attribute utilities. *Decision Sciences* **5**, 430–458.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association* **89**, 309–319.

Inhelder, B. and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York: Basic Books.

Jenkins, H. M. and Ward, W. C. (1965). The judgement of contingency between responses and outcomes. *Psychological Monographs* **79** (1, Whole No. 594).

Johnson, E. J., Hershey, J., Meszaros, J. and Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty* **7**, 35–51.

Kadane, J. B. (1980). Predictive and structural methods for eliciting prior distributions. In *Bayesian Analysis in Econometrics and Statistics: Essays in honor of Harold Jeffreys*, A. Zellner (ed.), 89–93. North Holland Publishing Company, Amsterdam.

Kadane, J. B. (ed.) (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*, J. Wiley & Sons.

Kadane, J. B., Chan, N. H. and Wolfson, L. J. (1996). Priors for unit root models. *Journal of Econometrics* **75**, 99–111.

Kadane, J. B. and Chuang, D. T. (1978). Stable decision problems. *Annals of Statistics* **6**,1095–

1110.

Kadane, J. B., Dickey, J., Winkler, R., Smith, W. and Peters, S. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* **75**, 845–854.

Kadane, J. B. and Schum, D. A. (1996). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. J. Wiley and Sons, New York.

Kadane, J. B. and Winkler, R. L. (1988). Separating Probability Elicitation from Utilities. *Journal of the American Statistical Association* **83**, 357–363.

Kadane, J. B. and Wolfson, L. J. (1998). Experiences in elicitation. *The Statistician* **47**, 1–20 (with discussion, pp 55–68).

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review* **80**, 237–251.

Keren, G. (1991). Calibration and probability judgements: conceptual and methodological issues. *Acta Psychologica* **77**, 217–273.

Koehler, J. J. (1996). The base rate fallacy reconsidered: normative, descriptive and methodological challenges. *Behavioral and Brain Sciences* **19**, 1–17.

Koriat, A., Lichtenstein, S. and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and Memory* **6**, No. 2.

Kunda, Z. and Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology* **18**, 195–224.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology* **73**, 498–502.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical*

Society B **57**, 247–262.

Lehrer, K. (1976). When rational disagreement is impossible, *Nous* **10**, 327–332.

Lichenstein, S., Fischhoff, B. and Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In *Judgement and Uncertainty: Heuristics and Biases*, D. Kahneman et al. (eds.), 306–334. Cambridge: Cambridge University Press.

Lichenstein, S. and Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance* **26**, 149–171.

Lichtenstein, S. and Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science* **9**, 563-564.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In *Bayesian Statistics 2*, J. M. Bernardo et al. (eds.), 375–390. Amsterdam: North Holland Publishing Company.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979). On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. A* **142**, 146–180.

Madansky, A. (1978). Externally Bayesian groups. Unpublished manuscript, University of Chicago.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* **22**, 1087–1096.

McConway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association* **76**, 410–414.

Meyer, M.A. and Booker, J.M. (2001). Eliciting and Analyzing Expert Judgment: A Practical Guide, *ASA-Society of Industrial and Applied Mathematics*, Philadelphia, PA.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: Cambridge University Press.

Morris, P. A. (1974). Decision analysis expert use. *Management Science*, **20**, 1233–1241.

Mosteller, F. and Yountz, C. (1990). Quantifying probabilistic expressions (with discussion). *Statistical Science* **5**, 2–34.

Murphy, A. H. and Winkler, R. L. (1974) Credible interval temperature forecasting: some experimental results. *Monthly Weather Review* **102**, 784–794.

Nash, H. (1964). The judgement of linear proportions. *American Journal of Psychology* **77**, 480–484.

Nisbett, R. E., Borgida, E., Crandall, R. and Reed, H. (1976). Popular induction: information is not necessarily informative. *Cognition and Social Behaviour*. Potamic, Md. : Lawrence Erlbaum Associates.

Oakley, J. E. and O'Hagan, A. (2002). Uncertainty in prior elicitations: a nonparametric approach. *Research Report No. 521/02* Department of Probability and Statistics, University of Sheffield.

Oberkampf, W. L., Helton, J. C., Joslyn, C. A., Wojtkiewicz. S. F. and Ferson, S. (2004). Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety* (in press).

O'Hagan, A. (1988). *Probability: Methods and Measurement*. Chapman and Hall, London.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician* **47**, 21–35 (with discussion, pp 55–68).

O'Hagan, A. and Oakley, J. E. (2004). Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety* (in press).

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics Volume 2B, Bayesian Inference* (2nd edition). Edward Arnold, London.

Oman, S. D. (1985). Specifying a prior distribution in structured regression problems. *Journal of the American Statistical Association* **80**, 190–195.

Peterson, C. R. and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin* **68**, 29–46.

Peterson, C. R. and Miller, A. (1964). Mode, median and mean as optimal strategies. *Journal of Experimental Psychology* **68**, 363–367.

Peterson, C. R., Snapper, K. J. and Murphy, A. H. (1972). Credible interval temperature forecasts. *Bulletin of the American Meteorological Society* **53**, 966–970.

Phillips, L. D. (1999). Group elicitation of probability distributions: Are many heads better than one? In J. Shanteau et al. (eds.), *Decision Science and Technology: Reflections on the Contributions of Ward Edwards* (pp. 313–330). Norwell, MA: Kluwer Academic Publishers.

Pill, J. (1971). The Delphi method: Substance, context, a critique, and an annotated bibliography. *Socio-Econ. Plan. Sci.* **5**, 57–71.

Pitz, G. F. (1965). Response variables in the estimation of relative frequency. *Perceptual and Motor Skills* **21**, 867–873.

Pitz, G. F. (1966). The sequential judgement of proportion. *Psychonomic Science* **4**, 397–398.

Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choice Under Uncertainty.* Addison-Wesley: Reading, Mass.

Schaefer, R. E. and Borcherding, K. (1973). The assessment of subjective probability distributions: a training experiment. *Acta Psychologics* **37**, 117–129.

Seaver, D. A., von Winterfeldt, D. and Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance* **21**,

379–391.

Seidenfeld, T., Schervish, M. J. and Kadane, J. B. (1989). On the shared preferences of two decision makers. *Journal of Philosophy* **86**, 225–244.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure type, and task. *Journal of Experimental Psychology* **61**, 430–436.

Simpson, W. and Voss, J. F. (1961). Psychophysical judgements of probabilistic stimulus sequences. *Journal of Experimental Psychology* **62**, 416–422.

Singpurwalla, N. D. and Song, M. S. (1987). The analysis of Weibull lifetime data incorporating expert opinion. In *Probability and Bayesian Statistics*, R. Viertl (ed.), 431–442. New York: Plenum.

Slovic, P. (1972). From Shakespeare to Simon: speculations – and some evidence – about man's ability to process information. *Oregon Research Institute Monograph*, Vol. 12, No. 2.

Slovic, P. and Fischhoff, B. (1977). On the psychology of experimental surprises. *Organizational Behavior and Human Performance* **3**, 544–551.

Slovic, P. and Lichtenstein, S. C. (1968). The relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology Monograph Supplement* **78**, No. 3, Part 2.

Slovic, P. and Lichtenstein, S. C. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior and Human Performance* **6**, 649–744.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology* **4**, 165–173.

Spencer, J. (1961). Estimating averages. *Ergonomics* **4**, 317–328.

Spencer, J. (1963). A further study of estimating averages. *Ergonomics* **6**, 255-265.

Spetzler, C. S. and Stael von Holstein, C-A. S. (1975). Probability encoding in decision analysis. *Management Science* **22**, 340–358.

Stael von Holstein, C.-A. S. (1971). an experiment in probabilistic weather forecasting. *Journal of Applied Meteorology* **10**, 635–645.

Stevens, S. S. and Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology* **54**, 377–411.

Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society B* **36**, 148–159.

Tversky, A. and Kahneman, D. (1971). The belief in the law of small numbers. *Psychological Bulletin* **76**, 105–110.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* **5**, 207–232.

Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science* **185**, 1124–1131.

Tversky, A. and Koehker, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review* **10**, 547–567.

Wallsten, T. S. and Budescu, D. V. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Science* **29**, 151–173.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R. and Forsyth, B. (1986). Measuring vague meanings of probability terms. *Journal of Experimental Psychology: General* **115**, 348–365.

Ward, W. C. and Jenkins, H. M. (1965). The display of information and the judgement of contingency. *Canadian Journal of Psychology* **19**, 231–241.

Wiggins, N. and Hoffman, P. J. (1968). Three models of clinical judgement. *Journal of Abnormal Psychology* **73**, 70–77.

Windschitl, P. D. and Wells, G. L. (1996). Measuring psychological uncertainty: verbal versus numeric methods. *Journal of Experimental Psychology:Applied* **2**, 343–364.

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association* **62**, 776–800.

Winkler, R. L. (1972). *Introduction to Bayesian Inference and Decision.* New York: Holt, Rinehart & Winston.

Winkler, R. L. (1980). Prior information, predictive distributions, and Bayesian model building. In *Bayesian Analysis in Econometrics and Statistics: Essays in honor of Harold Jeffreys*, A. Zellner (ed.), 95–109. North Holland Publishing Company, Amsterdam.

Wolfson, L. J. (1995). Elicitation of priors and utilities for Bayesian analysis. PhD Thesis. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Youssef, Z. I. and Peterson, C. R. (1973). Intuitive cascaded inferences. *Organizational Behavior and Human Performance* **10**, 349–358.

Zellner, A. (1972). On assessing informative prior distributions for regression coefficients. Unpublished mimeo.