Eliciting Probability Distributions

Jeremy E. Oakley

March 4, 2010

1 Introduction

In this chapter we discuss the process of eliciting an expert's probability distribution: extracting an expert's beliefs about the likely values of some unknown quantity of interest, denoted by θ, and representing those beliefs with a probability distribution. We consider the case of a scalar continuous θ only, so that a univariate density function is used to represent the expert's beliefs. The quantity θ can be, for instance, a hyperparameter needed in a Bayesian analysis, or any other parameter of specific interest.

We suppose that the elicitation involves three individuals. Firstly, there is a decision-maker who needs to know θ for a decision problem. Secondly, there is an expert who has knowledge about θ, with the decision-maker wishing to get the expert's beliefs about θ in the form of a probability distribution. Finally, there is a facilitator, who has expertise in the process of probability elicitation, and will get a probability distribution from the expert. Two or more of these individuals could be the same person. In Bayesian decision theory, the decision-maker should use his or her own probability distribution for θ in the decision-making problem. In practice, however, the decision-maker may be willing to use a distribution specified directly by an expert, if the expert is judged to have considerably more relevant knowledge, and is trusted to be impartial. (Lindley et al. (1979) discuss the problem of how to assess one's own probability distribution given probability judgements from another, but we do not consider this here).

The facilitator and expert are likely to be different people, and to simplify exposition, we suppose that the facilitator is male and the expert is female. In many cases, prior to the elicitation, the expert will not have considered her beliefs about θ probabilistically, and will be unfamiliar with the process of elicitation. Hence the facilitator has an important role in helping the expert make probabilistic judgements.

In sections 2 and 3 we discuss some general principles that the facilitator should consider when preparing for and conducting an elicitation session. In the remaining sections, we work through the process of elicitation, starting from the planning stages.

2 Eliciting distributions versus estimating proportions

The literature on probability elicitation covers two distinct but related tasks: using a probability distribution to represent subjective uncertainty about some quantity (the subject of this chapter), and estimating a proportion. The distinction is not always made clear, but it is important that the distinction is understood by all concerned in the elicitation process.

The following is an example of the first task.

> Let θ be the date in which the first human settlers arrived in Europe. What probability distribution represents your uncertainty about θ?

Clearly, we are using probability to describe *epistemic* uncertainty: uncertainty due to lack of knowledge, rather than to describe *aleatory* uncertainty: uncertainty due to randomness. There is no true probability distribution, and two experts can legitimately specify different distributions, as one may have more relevant knowledge than the other, or they may have disagreements regarding the available evidence. An expert may underestimate or overestimate θ, but it does not make sense to say that she has underestimated or overestimated *her probability* regarding θ, for example $P(\theta < 500,000BC)$.

An example of the second task would be

> Estimate *the* probability of an adult male being involved in a road accident this year?

Here, it would be assumed that the adult male is randomly selected from some population, and the true probability is defined to be the proportion of adult males in the population who are involved in road accidents in the given year. Although the word "probability" is typically used in this context in the literature, from here on we refer to this task as estimating a proportion, to keep the distinction between the two tasks clear. There has been a large volume of research investigating how people estimate proportions, and factors that can lead to biased estimates. For example, Tversky and Kahneman (1974) discuss the "availability heuristic", in which an individual estimates a proportion based on the ease of which relevant instances come to mind. In the example, this could lead to an inflated estimate if the expert has recently witnessed a severe road accident.

In a decision-making or inference problem where uncertainty is important, it will almost always be the first type of elicitation task that is required and not the second, because the second task does not involve any consideration of uncertainty. In the road accident example, if the expert estimates the proportion to be 0.01, this does not imply that she is *certain* that 1% of all adult males will be involved in road accidents. If we want to consider the expert's uncertainty, the question should be re-phrased as

> Let θ be the proportion of adult males in the population who are involved in road accidents in the given year. What probability distribution represents your uncertainty about θ?

Some findings reported in the literature are clearly relevent in both tasks. For example, the availability heuristic has implications for both the task of estimating a proportion, and in providing a distribution to represent one's uncertainty about a proportion. Other results do not so obviously transfer from one context to the other. For example, Gigerenzer (1996) reports an experiment in which physicians could estimate proportions more accurately when presented with information in a frequency format; considering numbers of patients out of one hundred, rather than simply a proportion of patients as a number between zero and one. However, the frequency format does not always make sense in the elicitating distributions context (e.g. we cannot consider one hundred Europes), and even when it does, it should be used with caution. In the road accident example, we might ask the expert

> Consider a random sample of 1000 adult males. What is your uncertainty about the number X of those males involved in road accidents?

A similar example from finance could be the following.

> Consider a random sample of 1000 BB-rated companies. What is your uncertainty about the number X of firms going bankrupt within the next year?

Phrasing questions in this way introduces two sources of uncertainty: epistemic uncertainty due to lake of knowledge about the population proportion θ, and aleatory uncertainty due to randomly sampling 1000 adults and companies, respectively. These two sources of uncertainty can be dealt with separately by first considering a distribution for θ, and then supposing that X given θ has a binomial distribution, but this would defeat the purpose of using a frequency format.

2.1 Measuring the quality of elicited distributions and calibration

If an expert is estimating a proportion, it is straightforward to measure the error in the estimate, given the true value, and make some assessment regarding the quality of the estimate. It is not so straightforward to compare an elicited distribution with the corresponding 'true' value, and determine whether the elicited distribution was 'good' or not. If, for example, the observed θ happens to be the 99th percentile of the distribution function elicited by the expert, this does not imply that expert made a 'bad' probability judgement. The expert may have considered her uncertainty sensibly in this instance, with θ taking a value in the tails of her distribution, as should occasionally happen.

One way to compare an elicited distribution with the observed true value is to use a scoring rule. Various scoring rules are discussed in Matheson and Winkler (1976), for example, the quadratic score $2f(\theta^*) - \int_{-\infty}^{\infty}\{f(\theta)\}^2 d\theta$, where $f(\cdot)$ is the elicited density function, and θ^* is the observed true value of the variable of interest. The absolute value of a particular score may not be easily interpretable, but differences in scores can be used to monitor improvements in an expert's ability to make probability assessments, or compare the performance of different elicitation methods.

A more intuitive measure that has been used to quantify the quality of elicited judgements is that of calibration. To measure an expert's calibration, she must provide a large number of probability judgements for unrelated quantities (and so calibration cannot be used to measure the quality of a *single* elicited distribution). Calibration is concerned specifically with probabilities rather than distributions, and so given a complete elicited distribution one would first extract a single probability, e.g. $P(a < \theta < b) = 0.95$. We then examine all the events in which the expert gave a probability of p of occurring, and calculate the proportion of these events which actually did occur. If the proportion is equal to p, for all values of p, then we say that the expert is perfectly calibrated. Although calibration can only be assessed over the course of a large number of probability assessments, it can be helpful for both facilitator and expert to imagine the elicitation as one of a series of elicitations, with good calibration being the aim in the long run.

There have been many studies investigating calibration, often reporting *overconfidence* in which, say, only 75% of events assigned probabilities of 95% occurred (see for example Alpert and Raiffa, 1982). Many of these studies used undergraduate students as the probability assessors (who, perhaps importantly, do not interact with a facilitator), and there has been some discussion of whether the results necessarily apply to experts. Reviews of the literature concentrating on expert assessment of probabilities are given in (Morgan and Henrion, 1990, pp. 128-136) and (O'Hagan et al., 2006, pp. 72-77), where the findings have been mixed. One example of very good calibration is the study of weather forecasters reported in Murphy and Winkler (1977). Arguably, an important feature in this case is that the weather forecasters were repeatedly making probability assessments, with prompt feedback following each assessment. A striking example of overconfidence was given by Ben-David et al. (2007). They found that financial executives are miscalibrated: realized market returns are within the executives' 80% probability intervals only 38 % of the time.

3 The psychology of probability assessment

The topic of elicitation is of obvious interest to psychologists given the central role of human judgement. There is a substantial literature on the subject, and recent reviews are given in O'Hagan et al. (2006) and Kynn (2008) (which facilitators are recommended to study). As discussed previously, it is important to consider the distinction between the two types of elicitation task (representing subjective uncertainty, and estimating a proportion) when studying this literature. We now summarise some important findings, although this section is not intended to be a comprehensive review.

An important observation is made by Winkler (1967):

> "The [expert] has no built-in prior distribution that is there for the taking. That is, there is no 'true' prior distribution. Rather, the [expert] has certain prior knowledge which is not easy to express quantitatively without careful thought. An elicitation technique used by the [facilitator] 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."

Hence facilitators should be aware that asking for the same judgement in two different, but mathematically equivalent ways may produce two different answers. One illustration of this is given in a calibration experiment reported in Soll and Klayman (2004). Participants were asked to provide 80% intervals for a number of quantities uncertain to them (for example, the date of Charles Dickens' birth). Participants in one group were asked directly for an 80% interval. Participants in a second group were asked first for their 10th percentiles, and then for their 90th percentiles. Overconfidence was observed in both groups, but to a lesser extent in the second group. The authors conjecture that in the second group, being prompted directly to first consider how small the uncertain quantity might be, and then how large it might be encouraged more thought about uncertainty and hence less overconfidence.

A major body of work in this area is the “heuristics and biases” research programme of Tversky and Kahneman (see for example Tversky and Kahneman, 1973): an investigation into various ‘rules of thumb’ for making probability judgements, and the biases that such rules can cause. (There has since been some debate regarding their findings, which is reviewed in Kynn, 2008). One such bias that we consider here is the “anchoring and adjustment effect”. A typical experiment would involve asking a participant first to consider whether some quantity that is uncertain to them, for example, the population of South Africa, is greater or less than some value X. The participant would then be asked to estimate the uncertain quantity. The participant may then ‘anchor’ on the value of X, and ‘adjust’ to give an estimate, with the adjustment not sufficiently far. In particular, using two different choices of X for two groups can result in significantly different estimates.

Anchoring and adjustment has three implications for expert elicitation. Firstly, if the facilitator proposes a value for θ, the expert may anchor on this value when considering her own beliefs. Secondly, when when considering uncertainty about θ, the expert may anchor on her ‘best estimate’ of θ, and adjust to obtain some percentile, perhaps insufficiently, leading to overconfidence. Thirdly, when the facilitator proposes a *distribution* for θ, the expert may anchor on this choice of distribution, possibly being too willing to accept the proposed distribution.

Another important observation that has been made is that the probability an individual gives to an event can be affected by the description of that event. Fischhoff et al. (1978) describe a series of experiments in which participants were given a list of possible causes of a car failing to start, and were asked to assign probabilities to each cause, in the event of a starting failure. They found that, even with expert mechanics, probabilities given for a particular cause, eg “battery”, were affected by what other causes were listed (the probability assigned to battery could be reduced by adding other causes to the list). We return to this example later in this chapter.

4 Planning an elicitation session

O’Hagan et al. (2006) give a model for the elicitation process consisting of five stages:

1. background and preparation;
2. identifying and recruiting the expert(s);
3. motivation and training the expert(s);
4. structuring and decomposition (typically deciding precisely what variables should be elicited, and how to elicit joint distributions in the multivariate case);

5. the elicitation itself.

This model might be used in a scenario where an expert is only available for a single meeting with the facilitator, with the preparation done prior to the meeting. However, this is not ideal, as it can be hard to prepare for an elicitation without subject matter expertise.

Having identified a need to obtain a distribution for θ, it is not always simply a matter of asking an expert to make judgements about θ directly; we may instead construct a distribution for θ from an elicited distribution of some other variable(s). The facilitator cannot always anticipate this in advance of his first meeting with the expert, hence earlier expert involvement can be desirable. It is worth noting the comment in Berger (2006) that

> "One only has limited time to elicit models and priors from the experts in a problem, and usually it is most efficient to use the available expert time for modeling, not for prior elicitation. Indeed, in the model construction phase of an analysis, it is usually quite counterproductive to perform subjective prior about parameters of models, since the models will likely not even be those used at the end."

Here we consider a slightly different model of elicitation. We suppose, having identified a need to obtain a distribution for θ, the decision-maker first recruits an expert (and facilitator), under the assumption that expert judgement is likely to be required. The facilitator then explains the elicitation process to the expert (the motivation and training stage), so that the expert is able to understand how elicitation works and when it is likely to be appropriate. The facilitator and expert should then discuss the available evidence, and decide precisely how/if formal expert elicitation should be used. It will then be necessary for the facilitator and expert to structure the elicitation problem, deciding precisely what variables to elicit, before conducting the elicitation session itself. We proceed through these stages in turn.

5 Motivation and training

First, the objectives of the analysis/decision-problem should be made clear to the expert, so that she understands how her judgements will be used. The facilitator can then proceed with the training, which can be divided into three parts: a general discussion of subjective probability distributions and their elicitation; a discussion of relevant known biases from the psychology literature; a practise elicitation exercise.

In the general discussion, the facilitator should cover the following.

1. The expert should understand the difference between representing subjective uncertainty with a probability distribution, and 'estimating a probability' (as discussed in section 2). This is particularly important if θ is itself a proportion (such as the percentage default rate in a fairly homogenous credit risk portfolio), as it is easy to confuse 'the probability' of an event with uncertainty about the proportion of cases in which an event occurs/is true. In any case, the expert should be clear that the objective is to get a probability distribution that represents her uncertainty.

2. The facilitator should make it clear that, when eliciting a distribution, he is *not* trying to obtain an artificially precise estimate of θ. His aim is to represent the expert's uncertainty as faithfully as possible. If little is known about θ, then the elicited density should reflect this. Alternatively, if the expert has good reasons for 'ruling out' certain parameter values, this too should be represented appropriately. It can be helpful to discuss the concept of calibration here, for example, if the expert were to provide a large number of independent 50% intervals for different quantities, then only 50% of them should contain the true values.

3. It can be helpful to give an example in which uncertainty is purely due to lack of knowledge, rather than randomness. For example, let θ be the shortest distance in miles by road between Paris and Berlin. The facilitator could display the $\mathcal{N}(200, 1)$ and the $\mathcal{U}[0, 10000]$[1] density functions, explain the judgements about θ that these represent, and highlight the fact that these are (likely to be) poor representations of anyone's beliefs. This can help motivate the idea of finding a suitable distribution to *faithfully* represent an individual's beliefs. A similar situation can be thought of when the facilitator wants to ask an economist about the Gross Domestic Product of the USA in 2012.

4. The facilitator should highlight the undesirability of the alternatives to eliciting a probability distribution that might be used. One alternative would obviously be just to fix the uncertain parameter at some estimated value, ignore any uncertainty, and proceed with the analysis. Another would be to use non-expert judgement. In particular, the expert should appreciate the value of using her elicited beliefs rather than a layperson's in the subsequent analysis.

5. The issue of eliciting expert judgement versus collecting more data should be discussed. Collecting data may simply not be feasible within the timeframe of the analysis. However, prior knowledge can play an important role when planning data collection, for example, in assessing whether a particular study would result in a

[1] We use the notation $\mathcal{N}(200, 1)$ to represent the normal distribution with mean 200 and variance 1 and $\mathcal{U}[0, 10000]$ to represent the uniform distribution over the interval $[0, 10000]$.

useful reduction in uncertainty (see Stevenson et al., 2009, for a case study), or investigating which uncertain variables are the most important for decision-making, and would be a priority for further investigation (Oakley, 2009). In strategic risk management, expert judgement might provide the decision maker (i.e. the senior management) with valuable information of whether a potential acquisition has the desired effect on the overall risk-return profile of the company.

6. The expert may find it helpful to consider 'reacting' to proposed values of θ. An expert may find she can quickly judge another person's estimate to be too high or too low, which can then encourage her to think what values of θ she would find acceptable. This sort of reasoning can help lead on towards the sorts of probabilistic judgements required for elicitation. (We recommend that the facilitator uses this approach in practice elicitation exercises only. Once the expert is comfortable with the elicitation process, the facilitator should avoid proposing values of θ, in case of anchoring effects).

The facilitator should then discuss relevant biases that have been reported in the psychology literature, such as anchoring effects and overconfidence. Other biases, for example availability, may be relevant, depending on the context.

Finally, the facilitator should conduct a practice elicitation, in which the true value of the variable is known to the facilitator, but not to the expert. If possible, the variable should be chosen from the same subject domain as θ, and the more practice exercises that can be done, with the facilitator giving the expert feedback on her performance, the better.

6 Identifying the role of elicitation

Having identified a need for a distribution for θ, and assuming that all concerned understand the elicitation process, the next step is to consider what evidence is available, and what the role of expert judgement will be in constructing the distribution. As an example, suppose the decision-maker wishes to consider uncertainty about the proportion θ of cannabis users within some population. We denote the information relevant to θ by D, so that the decision-maker wishes to construct $f(\theta|D)$. We now consider three scenarios (which could occur simultaneously).

6.1 Strong, directly relevant evidence

Suppose first that D consists of a large sample survey, in which N members of the population are randomly sampled, and (truthfully) state whether they are cannabis users or not. In this case, Bayes' theorem could be used to obtain $f(\theta|D)$:

$$f(\theta|D) = \frac{f(\theta)f(D|\theta)}{\int f(\theta)f(D|\theta)d\theta}.$$

(See Chapter 1 of this book for a general introduction to Bayesian statistics). A binomial likelihood function would be suitable here, so that $D|\theta \sim Binomial(N, \theta)$. Technically the prior density $f(\theta)$ should represent the decision-maker's prior beliefs about θ, but the decision-maker may prefer to ask a trusted expert to specify $f(\theta)$. Should expert elicitation be used to obtain $f(\theta)$?

One problem is that the expert may already know the information D. She would then have to consider what she would believe about θ *if she did not know* D. It is debateable whether she would be able to make good judgements in such circumstances. Specifically, there would be a risk of *hindsight bias*, in which she believes she would have predicted the observed data (Fischhoff, 1975). She may also judge the elicitation process to lack credibility if she believes her prior should have little weight in comparison to the data D.

In this scenario, the decision-maker should first explore alternatives to formally elicited priors. Noninformative priors could be used (see Kass and Wasserman, 1996, for a review), or the decision-maker could consider 'sceptical' or 'enthusiastic' priors of the sort discussed in Spiegelhalter et al. (2004, pp. 158-161). If decisions or inferences are robust to a range of candidate priors, then formal elicitation is unlikely to be worthwhile.

6.2 Indirectly relevant evidence

Alternatively, suppose that expert is aware of the data D, but does not believe it is directly suitable for making inferences about θ. This scenario is discussed in Turner et al. (2008) who give a case-study in using expert judgement to adjust for biases in individual studies when combining the studies in a meta-analysis. They classify biases under two headings: "internal biases", in which flaws in a study design can lead to biased estimates, and "external biases", in which a study was designed to estimate a different, but related quantity.

As an example of an internal bias, returning to the cannabis example, the expert may believe that not all participants may respond truthfully. We could now define ϕ to be the proportion of all cannabis users that would respond truthfully (given the survey questions used). If the judgement is made that $D|\theta, \phi \sim Binomial(N, \theta\phi)$ then we have

$$f(\theta|D) \propto f(\theta)f(D|\theta) = f(\theta)\int f(D|\theta, \phi)f(\phi)d\phi,$$

where the facilitator could elicit a proper prior for ϕ, and use a noninformative prior for θ.

As an example of an external bias, the decision-maker may be interested in the proportion of cannabis smokers in the UK adult population, but only have data from a survey of US adults. In this case, we could elicit expert judgement about the difference between the two population proportions, and then use the data about the US proportion to update beliefs about the UK proportion.

6.3 Expert opinion only

A third scenario is that the information D is not in the form in which one could specify a likelihood function. For example D may be an informal impression of the proportion of cigarette smokers in the population, together with the judgement that this is likely to exceed the proportion of cannabis smokers. In this case, the facilitator might directly elicit $f(\theta|D)$, thereby bypassing the formal process of Bayesian updating.

6.4 Structuring the elicitation problem

We define this stage to mean deciding precisely what variable(s) to elicit, and how to construct a (joint) distribution for the uncertain variable(s) of interest. Smith (1998) writes that in his experience

> "it is paramount to spend a significant proportion of my time eliciting structure: dependencies, functional relationships and the like."

We have already discussed one aspect of this, in considering what should be elicited given the available evidence. As another illustration, returning to the car example from Fischhoff et al. (1978), suppose θ is the proportion of all car journeys (say in a particular model of car) in which the car fails to start. Rather than asking for a judgements about θ directly, the facilitator may first ask the expert to produce a list of possible causes, and then elicit her beliefs about the frequency of each cause.

In the multivariate case, the facilitator and expert must discuss how to construct a joint distribution for the uncertain quantities. For example, suppose a joint distribution is required for two uncertain proportions θ_1 and θ_2, and the expert does not judge these variables to be independent (so that she would change her beliefs about θ_1 given the value of θ_2, and vice versa). She may judge θ_1 and the ratio θ_1/θ_2 to be independent, and so the facilitator would elicit independent distributions for θ_1 and θ_1/θ_2 which could then be combined to obtain the joint distribution of θ_1 and θ_2.

7 Eliciting a distribution

Having identified precisely what variable(s) to elicit, the facilitator can then proceed with eliciting a distribution. For any variable θ, this can be done using the following procedure.

1. The expert makes a small number of probabilistic judgements about θ.
2. The facilitator fits a suitable parametric probability distribution to the expert's judgements.
3. The facilitator reports features of the distribution back to the expert, and asks the expert whether the fitted distribution is an acceptable representation of her beliefs.
4. If the distribution is acceptable to the expert then the elicitation is concluded. Otherwise, the facilitator fits an alternative distribution, usually based on modified or additional probabilistic judgements from the expert.

There are various different types of judgement that the facilitator can ask for, leading to different elicitation methods. Two types of judgement we consider her are *fixed interval* and *variable interval* judgements. In the fixed interval method, the expert is asked for probabilities of the form $P(a < \theta < b)$. In the variable interval method, the expert is asked for quantiles; for example, the expert is asked for the value a such that $P(a < \theta) = 0.25$.

Abbas et al. (2008) describe a study using university students suggesting slight superiority of the fixed interval method. Garthwaite et al. (2007) did not find one method to be consistently better than the other (though they did note other substantial differences depending on whether the probability assessors were students or genuine subject-matter experts). Murphy and Winkler (1974) found the variable interval method to perform better, though theirs was a study of only four (genuine subject-matter) experts. We suggest, in the training stage of the process, experimenting with the different approaches described here, and using the one that the expert is most comfortable with.

7.1 SHELF

SHELF (the Sheffield Elicitation Framework) is a package of templates and software for conduction elicitations, and can be downloaded for free from `www.tonyohagan.co.uk/shelf`. The software is written in R (R Development Core Team, 2009), and includes interactive graphics routines developed using the `rpanel` library (Bowman and Crawford, 2008).

The templates are designed both to guide the facilitator through the process of conducting an elicitation, and to enable the facilitator to provide a thorough record of the elicitation session. For each template there is a blank version for use, and a version with commentary to help the facilitator.

7.2 The bisection method: a variable interval method

We first consider the case when θ is constrained to lie between 0 and 1, for example if θ is a proportion. One particular variable interval method that can be used is the bisection method, which is illustrated in Raiffa (1968, pp. 161-168).

1. The facilitator asks the expert to do the following.

    > Choose a value m, such that you judge the two intervals $[0, m]$ and $[m, 1]$ to have the same probability of containing θ.

    The facilitator can explain what is meant by "the same probability", by describing a gamble, in which, for a given m, the expert chooses one of the two intervals $[0, m]$ or $[m, 1]$, and receives a reward if θ lies in the chosen interval (but does not pay any penalty if θ lies in the other interval). If the expert judges the two intervals to have the same probability, then she should have no preference for one interval over the other in this gamble.

    The facilitator can help the expert make this judgement by proposing values of m, and asking her simply to consider which interval she judges to be more likely. For example the expert may judge $[0, 0.5]$ to be more probably than $[0.5, 1]$, but judge $[0, 0.25]$ to be less probable than $[0.25, 1]$, implying that m must be somewhere in the interval $[0.25, 0.5]$. Of course, it is preferable if the facilitator does not make any suggestions that might unduly influence the expert, but this sort of prompting can be used during the training phase of the elicitation, to help the expert consider how to choose her median value.

2. The facilitator now elicits the expert's lower quartile, l, by asking the expert to do the following.

> Divide the interval $[0, m]$ into two equally probable intervals $[0, l]$ and $[l, m]$.

Again, the same hypothetical gamble can be used to explain what is required. In the author's experience, the expert is likely to find this more difficult than choosing the median, and will need further help from the facilitator. The facilitator can help the expert by asking her to consider how uncertain she is about the value of θ; whether θ is very likely to be close to m, or whether it could be considerably lower. He can prompt her by asking her to consider whether she judges $[0, m/2]$ to be more or less probable than $[m/2, m]$ (to which the answer should almost always be that $[0, m/2]$ is less probable), and then asking her to consider whether $[0, m - \delta]$ is more or less probable than $[m - \delta, m]$, for some suitably small δ, leading the expert to consider a value for l in the interval $[m/2, m - \delta]$. Again, it is preferable to avoid leading the expert, but this can be useful at the training stage.

3. The facilitator now elicits the expert's upper quartile, u, noting the considerations in step 2, by asking her to do the following.

> Divide the interval $[m, 1]$ into two equally probable intervals $[m, u]$ and $[u, 1]$.

4. The facilitator now invites the expert to reflect on her choices and check for consistency, by asking

> Consider the four intervals $[0, l]$, $[l, m]$, $[m, u]$ and $[u, 1]$. Do you judge any one of them to be more probable than any other?

In theory, the expert should judge the four intervals to be equally probable, but it's quite possible that at this stage that, seeing her choices of l, m and u presented simultaneously, she may judge some intervals to be more probable than others. In this case, she is asked to modify her choices of l, m and u appropriately.

5. The facilitator now fits a parametric distribution to these judgements. An obvious choice given that θ is a proportion would be the beta distribution. The parameters can be fitted using a least squares approach: α and β are chosen to minimise

$$\{F(l; \alpha, \beta) - 0.25\}^2 + \{F(m; \alpha, \beta) - 0.5\}^2 + \{F(u; \alpha, \beta) - 0.75\}^2,$$

where $F(x; \alpha, \beta)$ is the cumulative distribution function of a beta random variable with parameters α and β:

$$F(x; \alpha, \beta) = \int_0^x \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta.$$

We denote the fitted distribution by $Beta(\hat{\alpha}, \hat{\beta})$. The minimisation cannot be done analytically, and so a suitable numerical optimisation algorithm should be used.

Routines for fitting distributions are available in the SHELF package, and example output is shown in Figure 1

The facilitator should compare the fitted quartiles (the lower quartile, median, and upper quartile of $Beta(\hat{\alpha}, \hat{\beta})$) with the elicited quartiles. Small discrepancies should be acceptable, as the expert will usually acknowledge some imprecision in her chosen quartiles. If the discrepancies are larger, it may be necessary to consider an alternative family of distributions (although we would expect the family of beta distributions to be sufficiently flexible in most cases). Difficulties in fitting can be caused by 'strange' positioning of the quartiles, e.g., by choosing the lower quartile too close to 0 relative to the median and upper quartile, but this would normally be picked up at stage 4.

6. The facilitator must now check whether the chosen distribution is an acceptable representation of the expert's beliefs, given that she has only provided him with three quartiles. This is known as the 'feedback' stage, and involves presenting the fitted distribution back to the expert, together with some additional summaries of the distribution. One option is to obtain the 0.33 and 0.66 quantiles from $Beta(\hat{\alpha}, \hat{\beta})$, which we denote by $\theta_{0.33}$ and $\theta_{0.66}$, so that $[0, \theta_{0.33}]$, $[\theta_{0.33}, \theta_{0.66}]$ and $[\theta_{0.66}, 1]$ are three equally probable intervals. The expert may wish to specify alternative values for these quantiles, and/or modify her original judgements, in which case the facilitator should refit the distribution using the least squares procedure.

7. Finally, the facilitator should discuss the tails of the fitted distribution with the expert, for example, by considering the 1st and 99th percentiles. Considering each tail in turn, the facilitator should ask questions such as "What event would have to happen for θ to be this small/large? How likely would such an event be?" If the expert is satisfied with the tails of the fitted distribution, then the elicitation is concluded. Otherwise, she may wish to revise her initial judgements, or the facilitator may need to fit an alternative distribution using the least squares procedure and incorporating the expert's proposed additional percentiles. The process is iterated until the expert is satisfied with the chosen fitted distribution. (The SHELF routines include a range of families of distributions for fitting).

7.2.1 The bisection method for unbounded distributions

The bisection method can be used for any type of continuous univariate distribution, in principle by asking the expert to divide the interval $(-\infty, \infty)$ into two equally likely intervals and so on. In practice, it may be necessary or helpful to consider finite bounds $\theta_{\min}$ and $\theta_{\max}$ first, with the assumption that $P(\theta < \theta_{\min}) = P(\theta > \theta_{\max}) = 0$. (This is required for the graphics routines in SHELF). Care is needed when choosing these values.

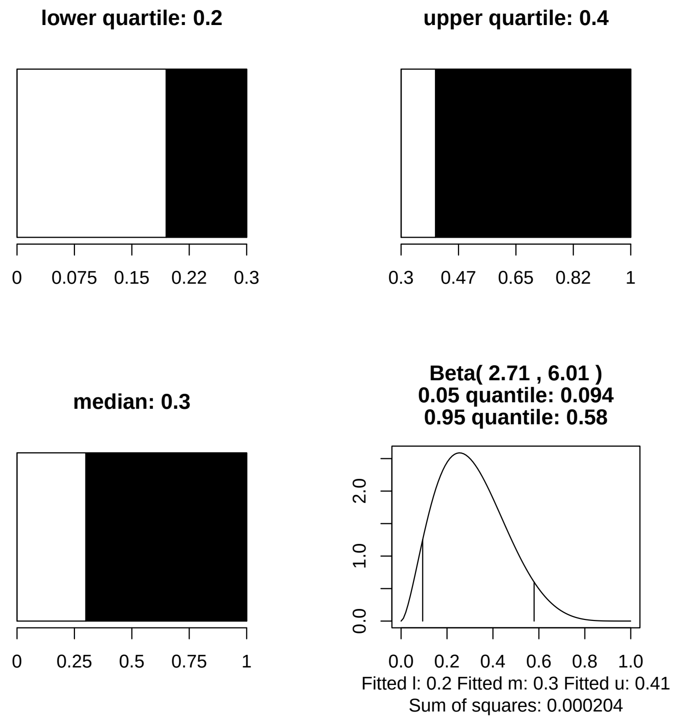

Figure 1: The bisection method. In each of the top and bottom left plots, the black and white regions represent regions of equal probability (corresponding to values $l = 0.2$, $m = 0.3$, and $u = 0.4$ as described in the text). The fitted beta distribution is shown in the bottom right plot.

Shorter intervals are preferable when using graphical methods to divide $[\theta_{\min}, \theta_{\max}]$, but this must be balanced against the judgement that the events $\{\theta < \theta_{\min}\}$ and $\{\theta > \theta_{\max}\}$ are impossible.

7.3 Fixed interval methods

In the fixed interval method, the facilitator presents intervals to the expert, and asks for her probabilities of θ lying in each interval. For example, in O'Hagan (1998) the facilitator asks the expert to provide lower and upper bounds $\theta_{\min}$ and $\theta_{\max}$, and a modal value r. The facilitator then asks the expert for her five probabilities: $p_1 = P(\theta_{\min} < \theta < r)$, $p_2 = P\{\theta_{\min} < \theta < (\theta_{\min} + r)/2\}$, $p_3 = P\{(r + \theta_{\max})/2 < \theta < \theta_{\max}\}$, $p_4 = P\{\theta_{\min} < \theta < (\theta_{\min} + 3r)/4\}$ and $p_5 = P\{(3r + \theta_{\max})/4 < \theta < \theta_{\max}\}$. (Technically, this involves a combination of variable and fixed interval methods, but the main emphasis is on fixed interval judgements).

Given a set of elicited probabilities, the facilitator can proceed following steps 5-7 in the bisection method procedure. A least squares approach can again be used to (numerically) fit a distribution. For example, if the facilitator wishes to fit a lognormal distribution to the elicited judgements, so that $\log\theta \sim \mathcal{N}(\mu, \sigma^2)$, he could do so by choosing μ and σ^2 to minimise

$$\left\{\Phi\left(\frac{\log r - \mu}{\sigma}\right) - \Phi\left(\frac{\log\theta_{\min} - \mu}{\sigma}\right) - p_1\right\}^2 + \dots$$
$$+\left\{\Phi\left(\frac{\log\theta_{\max} - \mu}{\sigma}\right) - \Phi\left(\frac{0.25\log(3r + \theta_{\max}) - \mu}{\sigma}\right) - p_5\right\}^2,$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

A fixed interval method is also implemented in SHELF, in which the expert is asked to provide $\theta_{\min}, \theta_{\max}$, her median m, and two probabilities $p_1 = P\{\theta_{\min} < \theta < (\theta_{\min}+2m)/3\}$ and $p_2 = P\{(2m + \theta_{\max})/3 < \theta < \theta_{\max}\}$. Example output is shown in Figure 2.

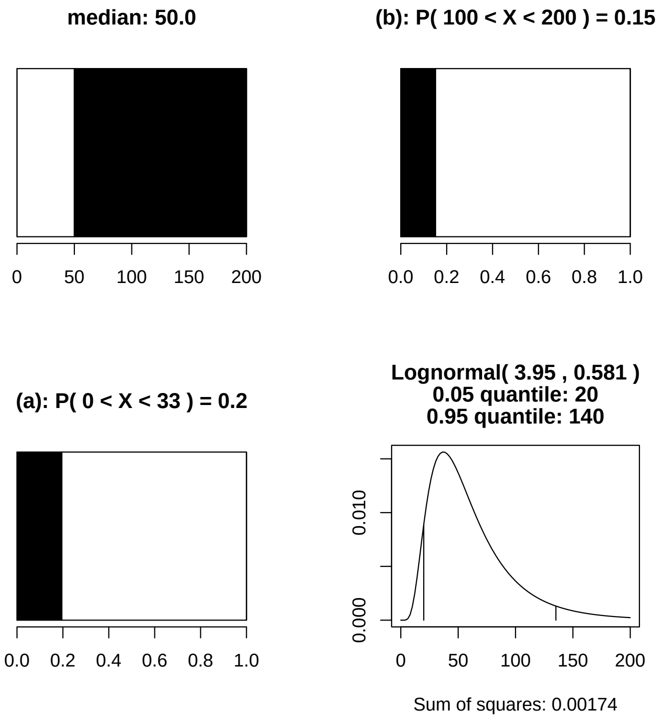

Figure 2: The fixed interval method in SHELF. (Note that the SHELF package uses X rather than θ to denote the uncertain quantity). The expert sets a range and median, and is then asked to provide two probabilities. The median is elicited as in the bisection approach. The top right and bottom left plots are now used to represent probabilities. In this example we have $\theta_{\min} = 0$, $\theta_{\max} = 200$ and $m = 50$.

7.4 The trial roulette method

The trial roulette method was proposed by Gore (1987), and is based on the fixed interval method. Some illustrations of this method can be found in Hughes (1991), Parmar et al.

(1994), Abrams et al. (1994), Parmar et al. (1996), Abrams and Dunn (1998) and Tan et al. (2003).

The sample space of θ is divided into m bins, and the expert is asked to distribute n chips amongst the bins, with the proportion of chips allocated to a particular bin representing the her probability of θ lying in that bin. If this is done graphically, the expert can see the shape of her distribution forming as she allocates the chips. Once the allocation has been done, the facilitator can fit a parametric distribution numerically as in the previous two methods, and should again proceed following steps 5-7 in the bisection method procedure.

This method is also implemented in SHELF, with example output in Figure 3. Here the expert is first asked to specify lower and upper bounds θ_{min} and θ_{max} for θ, with $[\theta_{\text{min}}, \theta_{\text{max}}]$ divided into ten bins. As before, a shorter interval $[\theta_{\text{min}}, \theta_{\text{max}}]$ is preferable, but this must be balanced against the judgement that the events $\{\theta < \theta_{\text{min}}\}$ and $\{\theta > \theta_{\text{max}}\}$ are impossible.

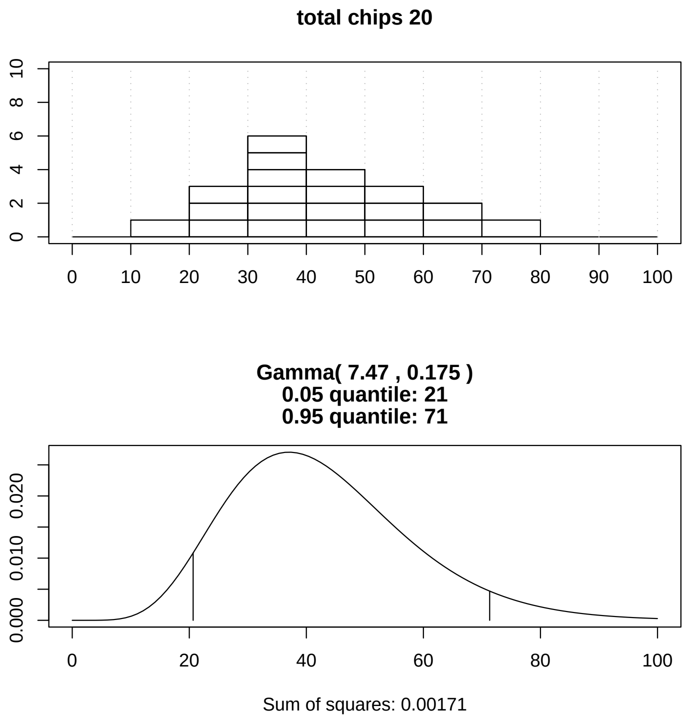

Figure 3: The trial roulette method. The expert allocates chips to bins, with the proportion of chips allocated to a particular bin interpreted as the expert's probability of θ lying in that bin. A distribution can then be fitted numerically as usual.

8 Discussion

In this chapter we have reviewed a general framework for eliciting univariate distributions, covering different techniques based on fixed interval and variable interval methods. We emphasise the following three points.

Firstly, preparation is important, and the expert should be involved as early as possible. Given the need to elicit a distribution for θ, it does not necessarily follow that this distribution will be obtained by asking for judgements about θ directly. Depending on the available evidence, or how the expert wishes to structure the problem, it may be better to elicit beliefs about some alternative quantity, from which the distribution of θ can be inferred.

Secondly, the choice of questions during the elicitation matter. The facilitator should not assume that two mathematically equivalent approaches will produce the same distribution, and should draw on lessons from the psychology literature where appropriate.

Finally, the facilitator should give the expert as much assistance as possible. This will involve giving suitable training, and if possible, exploring different elicitation methods with the expert to find which method she is most comfortable with. Having fitted a distribution to an initial set of judgements, the facilitator should offer feedback and invite further reflection from the expert, particularly regarding the tails of her distribution.

Many of the principles discussed here will apply in any elicitation context, though there are, or course, more complex problems involving multivariate distributions and/or multiple experts. These topics are discussed in chapters [XXX] in this book.

References

Abbas, A. E., Budescu, D. V., Yu, H.-T. and Haggerty, R. (2008). A comparison of two probability encoding methods: fixed probability vs. fixed variable values, *Decision Analysis*, **5**: 190–202.

Abrams, K., Ashby, D. and Errington, D. (1994). Simple Bayesian analysis in clinical trials - a tutorial, *Controlled Clinical Trials*, **15**: 349–59.

Abrams, K. R. and Dunn, J. A. (1998). Discussion on the papers on elicitation, *Journal Of The Royal Statistical Society Series D*, **47**: 60–61.

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors, in *Judgement and Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Ben-David, I., Graham, J. R. and Harvey, R. (2007). Managerial overconfidence and corporate policies, Tech. Rep. NBER Working Paper No. 13711.

Berger, J. (2006). The case for objective Bayesian analysis, *Bayesian Analysis*, **1**: 385–402.

Bowman, A. W. and Crawford, E. (2008). *R package* `rpanel`*: simple control panels (version 1.0-5)*, University of Glasgow, UK.

Fischhoff, B. (1975). Hindsight $\neq$ foresight: the effect of outcome knowledge on judgment under uncertainty, *Journal of Experimental Psychology: Human Perception and Performance*, **1**: 288–299.

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**: 330–344.

Garthwaite, P. H., Jenkinson, D. J., Rakow, T. and Wang, D. D. (2007). Comparison of fixed and variable interval methods for eliciting sibjective probability distributions, Tech. rep., University of New South Wales.

Gigerenzer, G. (1996). The psychology of good judgment: frequency formats and simple algorithms, *Medical Decision Making*, **16**: 273–280.

Gore, S. M. (1987). Biostatistics and the Medical Research council, *Medical Research Council News*.

Hughes, M. D. (1991). Practical reporting of bayesian analyses of clinical trials, *Drug information journal*, **25**: 381–93.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules, *J. Am. Statist. Assoc.*, **90**: 1343–70.

Kynn, M. (2008). The 'heuristics and biases' bias in expert elicitation, *J. R. Statist. Soc. A*, **171**: 239–264.

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

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

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

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

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature: Some preliminary results, *Applied Statistics*, **26**: 41–47.

Oakley, J. E. (2009). Decision-theoretic sensitivity analysis for complex computer models, *Technometrics*, **51**: 121–129.

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*, Chichester: Wiley.

Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. (1994). The CHART trials: Bayesian design and monitoring in practice, *Statistics in Medicine*, **13**: 1297–312.

Parmar, M. K. B., Ungerleider, R. S. and Simon, R. (1996). Assessing whether to perform a confirmatory randomised clinical trial, *Journal of the National Cancer Institute*, **88**: 1645–51.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

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

Smith, J. Q. (1998). Discussion note to the papers on 'Elicitation', *The Statistician*, **47**: 63–64.

Soll, J. B. and Klayman, J. (2004). Overconfidence in interval estimates, *Journal of Experimental Psychology: Learning, Memory and Cognition*, **30**: 299–314.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester: Wiley.

Stevenson, M. D., Oakley, J. E., Lloyd Jones, M., Brennan, A., Compston, J. E., McCloskey, E. and Selby, P. L. (2009). The cost-effectiveness of an rct to establish whether 5 or 10 years of bisphosphonate treatment is the better duration for women with a prior fracture, *Medical Decision Making*, **29**: 678–689.

Tan, S.-B., Chung, Y.-F., Tai, B.-C., Cheung, Y.-B. and Machin, D. (2003). Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma, *Controlled Clinical Trials*, **24**: 110–121.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2008). Bias modelling in evidence synthesis, *J. R. Statist. Soc. A*, **171**: 239–264.

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.

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