

# Nonparametric Prior Elicitation using the Roulette Method

Jeremy E. Oakley<sup>1</sup>, Alireza Daneshkhah<sup>2</sup> and Anthony O'Hagan<sup>1</sup>

<sup>1</sup> School of Mathematics and Statistics, University of Sheffield, UK and

<sup>2</sup>Department of Statistics, Shahid Chamran University, Iran

## Abstract

We consider the use of the roulette method for eliciting an expert's probability density function. In the roulette method, the expert provides probabilities of the uncertain quantity of interest lying in a particular 'bin' by allocating 'gaming chips' to that bin. This method can be appealing to some experts, given the graphical representation of their beliefs that it provides. Given the judgements made by the expert, we then quantify the uncertainty about their density function, given the fact the expert has only specified a limited number of probability judgements, and that these judgements are forced to be rounded. Uncertainty about the expert's density is quantified using a Gaussian process model, and we investigate the effect of the number of bins and chips used on this uncertainty.

KEY WORDS: Expert elicitation, Gaussian process, imprecise probabilities, trial roulette.

## 1 Introduction

We consider eliciting an expert's beliefs about some unknown continuous variable  $\theta$ . We suppose that the elicitation is conducted by a (male) facilitator, who interviews the (female) expert and identifies a density function  $f$  that represents her beliefs about  $\theta$ . He will help her as much as possible, for example by providing suitable training and discussing the various biases that can influence probability judgements.

An important practical issue in elicitation is regarding the type of probability judgements that the facilitator asks the expert to make. He will want to ask her questions that are as simple as possible to answer, lead to a faithful representation of her uncertainty, and avoid under- or overconfidence so that she is well calibrated.

Two commonly used elicitation methods are the fixed interval method, in which the expert states her probability of  $\theta$  lying in specified intervals, and the variable interval method, in which she makes quantile judgements. We are not aware of any conclusive evidence favouring one method over the other. Abbas et al. (2008) describe a study suggesting slight superiority of the fixed interval method, though the participants (university students) were making probability judgements without the support of a facilitator. 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 their study only had four (genuine subject-matter) experts.

In this paper we consider a type of fixed interval method known as (trial) roulette elicitation (Gore, 1987). The expert distributes  $n$  chips amongst  $m$  bins, with the proportion of chips allocated to a particular bin representing her probability of  $\theta$  lying in that bin. We refer to this proportion as her implied probability. Some illustrations can be found in Hughes (1991), Parmar et al. (1994), Abrams et al. (1994), Parmar et al. (1996), Abrams and Dunn (1998), Tan et al. (2003) and Johnson et al. (2010). This approach can be appealing to some experts, as they do not have to specify numerical probabilities directly, and the distributions of chips amongst bins gives an immediate graphical representation of their beliefs. Software for using the roulette method is provided as part of the SHELF package (Oakley and O'Hagan, 2010). The software is implemented in R (R Development Core Team, 2010) and includes interactive graphics using the rpanel package (Bowman and Crawford, 2008).

One feature of the roulette method is the lack of precision in the expert's implied probabilities. If the expert distributes a total of  $n$  chips, then her implied probabilities are forced to be multiples of  $1/n$ . Some experts may prefer smaller  $n$ , so that they only have to make coarse probability judgements, which will lower the precision. There is also the issue of where to locate the bins, in particular the endpoints. Garthwaite et al. (2007) observe that the choice of scale can have a substantial effect on an expert's judgements, though we do not consider this issue further here.

Given the allocation of chips to bins, the facilitator must choose a probability distribution to represent the expert's beliefs. He could simply use a piecewise uniform distribution, with the appropriate uniform density within each bin. This may not be desirable, as the expert may not wish to give zero probability to  $\theta$  lying outside the two end (non-empty) bins. Alternatively, he may fit some parametric distribution to her probabilities. But there are infinitely many distributions that he

could fit, all implying different judgements in the tails. Direct tail judgements could resolve this, but are not easy to make reliably. Alpert and Raiffa (1982) report in one experiment that judgements of extreme percentiles were poorly calibrated (98% intervals contained the true parameter values 53% of the time).

## 1.1 Aims and outline of this paper

We consider how to quantify the facilitator’s uncertainty about the expert’s density, given her roulette judgements. We also investigate how the choice of  $n$  affects the facilitator’s uncertainty. We do so using the nonparametric approach to elicitation proposed by Oakley and O’Hagan (2007) (hereafter O&O) and developed in Gosling et al. (2007) and Moala and O’Hagan (2010).

The idea in O&O is to treat elicitation as a Bayesian inference problem. The facilitator considers his prior beliefs about the expert’s density function  $f$ . He receives data from the expert, in the form of probability judgements about  $\theta$  implied by her allocation of chips to bins, and derives his posterior distribution for  $f$ . He can use his posterior distribution to provide an estimate of  $f$  and assess posterior uncertainty about  $f$ , thus determining whether the expert’s beliefs have been elicited in sufficient detail for the purpose at hand. As in O&O, we stress that the facilitator is only interested in the expert’s density function  $f$ ; he is not considering his own beliefs about  $\theta$  given judgements received from the expert.

In the next section we review the nonparametric elicitation scheme proposed by O&O. In section 3, we consider the extensions necessary for inference given the roulette data. A synthetic example is given in section 4, investigating the effect of different choices for the total and number of chips, and an application is given in section 5 involving expert beliefs about the effectiveness of a radiotherapy technique for lung cancer patients.

For reviews of expert elicitation in general we refer the reader to Garthwaite et al. (2005), O’Hagan et al. (2006), Daneshkhah and Oakley (2010) and Oakley (2010).

## 2 Nonparametric elicitation

The facilitator’s prior for  $f$  makes strong judgements about the smoothness of  $f$ , but relatively weak judgements about the value of  $f(\theta)$  for any  $\theta$ . As an illustration, consider the two density functions plotted in Figure 1. The facilitator would judge the top plot to be a more plausible representation of someone’s beliefs than the

bottom plot, given the large number of modes in the bottom plot, even without knowing what  $\theta$  represents.

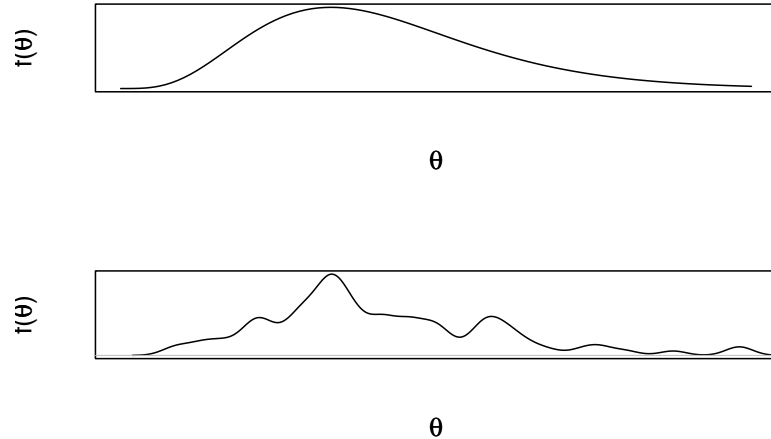


Figure 1: The facilitator expects the expert’s density function to be smooth, and so would judge the top plot to be a more plausible description of someone’s beliefs than the bottom plot.

The facilitator represents prior uncertainty about  $f$  using a Gaussian process: for any set  $\{\theta_1, \dots, \theta_n\}$  of values of  $\theta$ , his prior distribution for  $\{f(\theta_1), \dots, f(\theta_n)\}$  is multivariate normal. As  $f$  is a density function, two constraints are applied to the facilitator’s prior:  $\int_{-\infty}^{\infty} f(\theta)d\theta = 1$  and  $f(\theta) \geq 0$  for all  $\theta$ . The first constraint is applied as part of the data from the expert, and second constraint is applied (approximately) using simulation, which we discuss in later sections.

The facilitator specifies his mean and variance-covariance functions for  $f$ , hierarchically in terms of a vector  $\alpha$  of hyperparameters. His prior expectation of  $f(\theta)$  is some member  $g(\theta | u)$  of a suitable parametric family with parameters  $u$ , contained within  $\alpha$ , so

$$E\{f(\theta) | \alpha\} = g(\theta | u).$$

We follow O&O and choose  $g(\theta | u)$  to be a normal density function with mean  $m$  and variance  $v$ . Gosling et al. (2007) consider a  $t$  distribution as an alternative, but we do not consider this extension here.

His prior variance of  $f(\theta)$  depends on his expected size of  $f(\theta)$ , and this is modelled by a variance-covariance function with the scaled stationary form

$$\text{cov}\{f(\theta), f(\theta') | \alpha\} = g(\theta | u) g(\theta' | u) \sigma^2 c(\theta, \theta'),$$

where  $c(\theta, \theta')$  is a correlation function that takes the value 1 at  $\theta = \theta'$  and is a decreasing function of  $|\theta - \theta'|$ . The function  $c(.,.)$  must ensure that the prior

variance-covariance matrix of any set of observations of  $f(\cdot)$ , or functionals of  $f(\cdot)$ , is positive semidefinite. O&O use the function

$$c(\theta, \theta') = \exp\left\{-\frac{1}{2b}(\theta - \theta')^2\right\}.$$

The hyperparameters of this model are  $\alpha = (u, \sigma^2, b)$ . For  $u = (m, v)$  a noninformative prior  $\pi(m, v) \propto v^{-1}$  is chosen. For the parameter  $b$ , O&O consider prior beliefs about the ratio  $b^* = b/v$ , removing the dependence on the scale of  $\theta$ , and suggest  $\log b^* \sim N(0, 1)$ , as this allows for suitably varied behaviour in prior realisations of  $f$  in terms of smoothness and multimodality.

If the expert provides ‘noise-free’ probabilities (i.e. no rounding errors as induced by the roulette method), the facilitator can use a noninformative prior for  $\sigma^2$ , for example  $\pi(\sigma^2) \propto \sigma^{-2}$ . Given noise in the expert’s probabilities, the data cannot rule out  $f(\theta)$  being a normal density function, with  $\sigma^2 = 0$ , and an improper prior for  $\sigma^2$  can result in an improper posterior. We use the same prior as O&O in the case of noisy data:  $\sigma^{-2} \sim \text{Gamma}(1, 1)$ .

### 3 Roulette data

We represent the bins by the intervals  $[x_0, x_1), [x_1, x_2), \dots, [x_{k-1}, x_k)$  (with the possibility of  $x_0 = -\infty$  and  $x_k = \infty$ ). The expert allocates  $n_i$  chips to the  $i$ th bin,  $[x_{i-1}, x_i)$ , with  $\sum_{i=1}^k n_i = n$ . We define  $r_i = n_i/n$  to be the expert’s implied probability of  $\theta \in [x_i, x_{i+1})$ , with  $r = (r_1, \dots, r_k)$ .

#### 3.1 Imprecision

In section 2, we discussed the facilitator’s uncertainty about  $f$  due to the expert only providing a finite number of probability judgements. The roulette method introduces a second source of uncertainty: probabilities are only being stated in multiples of  $1/n$ . As other authors have commented (e.g. Good, 1980), no individual could be expected to provide arbitrarily precise probabilities (apart from in trivial circumstances). If  $n$  is small enough, the expert may at least judge her allocation of chips to bins to be ‘precise’, in that no other allocation would be acceptable to her. For example, in the extreme case  $n = 1$ , she would only have to consider which bin she thought was most likely, which she may be willing to state ‘precisely’. If  $n$  is not ‘small enough’, the allocation itself may be imprecise, before even considering rounding errors.

Various approaches have been proposed to deal with imprecision in probability assessments, such as robust Bayes methods (see for example Berger, 1990) and upper and lower probabilities (Walley, 1991). Here, we choose to model the expert's imprecision probabilistically. We suppose that the expert has a set of true, precise probabilities  $p = (p_1, \dots, p_k)$ , with  $p_i$  the expert's true probability of  $\{\theta \in [x_{i-1}, x_i]\}$ , and in section 3.2, we consider a simple model for  $r$  given  $p$ . We acknowledge the informality of such an approach, but argue that it still results in a useful mechanism for quantifying the facilitator's uncertainty about the expert's density (particularly in giving more weight to some density functions than others). Note that the assumption of a 'true' probability vector  $p$  is at least justifiable in the sense that, were the expert to specify probabilities in each bin directly, she may give probabilities that are different from  $r$ .

### 3.2 Sampling from the facilitator's posterior distribution

The facilitator's data consists of  $r$  together with the knowledge that  $\sum p_i = 1$  and  $f \geq 0$ , where we write  $f \geq 0$  to denote  $f(\theta) \geq 0 \forall \theta$ . The latter constraint implies  $p_i \geq 0 \forall i$ , which we denote by  $p > 0$ .

We sample from the facilitator's posterior  $\pi(f|r, \sum p_i = 1, f \geq 0, p > 0)$ . To deal with the constraint  $f \geq 0$ , we write

$$\begin{aligned} \pi\left(f|r, \sum p_i = 1, f \geq 0, p > 0\right) &\propto \pi\left(f, f \geq 0|r, \sum p_i = 1, p > 0\right) \\ &\propto \pi\left(f|r, \sum p_i = 1, p > 0\right) I(f \geq 0). \end{aligned}$$

Hence we first sample from  $\pi(f|r, \sum p_i = 1, p > 0)$ , ignoring the constraint  $f \geq 0$ , and then discard any negative  $f$ .

If the expert were to provide her true probabilities  $p$  rather than  $r$ , then it would be straightforward to derive the distribution of  $f|p, \alpha, \sigma^2$ . This is discussed in detail in O&O, and follows from the result that the joint distribution of  $f(\theta), p|\alpha, \sigma^2$  is multivariate normal. However, it is harder to derive the posterior distribution of  $f(\theta)|r, \alpha, \sigma^2$ , as it is not obvious how to construct the likelihood  $\pi(r|\alpha, f)$ . Instead, we treat  $p$  as a vector of nuisance parameters, and sample from  $\pi(f|r, \sum p_i = 1, p > 0)$  by obtaining joint samples from  $\pi(f, \alpha, p, \sigma^2|r, \sum p_i = 1, p > 0)$ , writing

$$\begin{aligned} \pi(f, \alpha, p, \sigma^2|r, \sum p_i = 1, p > 0) &= \pi(\alpha, p, \sigma^2|r, \sum p_i = 1, p > 0)\pi(f|\alpha, p, r, \sigma^2) \\ &= \pi(\alpha, p, \sigma^2|r, \sum p_i = 1, p > 0)\pi(f|\alpha, p, \sigma^2) \end{aligned}$$

Given a large sample of density functions, we can report estimates and pointwise bounds for  $f(\theta)$  for any  $\theta$  of interest. We use Gibbs sampling to sample from

$\pi(\alpha, p, \sigma^2 | r, \sum p_i = 1, p > 0)$ . We sample from  $\pi(\alpha, \sigma^2 | r, p) = \pi(\alpha, \sigma^2 | p)$  and from  $\pi(f | \alpha, p, \sigma^2)$  following the procedure described in O&O. We consider how to sample from  $\pi(p | r, \alpha, \sigma^2, \sum p_i = 1, p > 0)$  in the next section.

### 3.2.1 Sampling from $\pi(p | r, \alpha, \sigma^2, \sum p_i, p > 0)$

We have

$$\begin{aligned} \pi(p | r, \alpha, \sigma^2, \sum p_i = 1, p > 0) &\propto \pi(p, p > 0 | \alpha, \sigma^2, r, \sum p_i = 1) \\ &\propto \pi(p | \alpha, \sigma^2, r, \sum p_i = 1) I(p > 0) \\ &\propto \pi(p | \alpha, \sigma^2, \sum p_i = 1) \pi(r | p, \alpha, \sigma^2) I(p > 0). \end{aligned} \quad (1)$$

The distribution of  $p | \alpha, \sigma^2$  is multivariate normal:  $p | \alpha, \sigma^2 \sim N(H, \sigma^2 A)$  with element  $i$  of  $H$  given by

$$\begin{aligned} E(p_i | \alpha, \sigma^2) &= E_f \left\{ \int_{x_i}^{x_{i+1}} f(\theta) d\theta \middle| \alpha, \sigma^2 \right\} \\ &= \int_{x_i}^{x_{i+1}} g(\theta | u) d\theta, \end{aligned} \quad (2)$$

and element  $i, j$  of  $\sigma^2 A$  given by

$$\begin{aligned} Cov(p_i, p_j | \alpha, \sigma^2) &= Cov_f \left\{ \int_{x_i}^{x_{i+1}} \int_{x_j}^{x_{j+1}} f(\theta) f(\theta') d\theta d\theta' \middle| \alpha, \sigma^2 \right\} \\ &= \sigma^2 \int_{x_i}^{x_{i+1}} \int_{x_j}^{x_{j+1}} g(\theta | u) g(\theta' | u) c(\theta, \theta') d\theta d\theta'. \end{aligned} \quad (3)$$

Expressions for these integrals are given in O&O. The joint distribution of  $p, \sum p_i | \alpha, \sigma^2$  is also multivariate normal, so that  $p | \sum p_i = 1, \alpha, \sigma^2 \sim N(H^*, \sigma^2 A^*)$ , where  $H^*$  and  $A^*$  are straightforward to derive, as (2) and (3) give us the mean and variance of  $\sum p_i$  and covariance between  $\sum p_i$  and  $p_i$ .

We must now consider the likelihood function  $\pi(r | \alpha, p, \sigma^2)$ . We choose to give increasing weight to smaller values of  $\|r - p\|$ , and set  $\pi(r | \alpha, p, \sigma^2)$  to be (proportional to) a multivariate normal density function, with mean vector  $p$  and diagonal variance matrix  $\delta^2 I$ . Note that correlation in the rounding errors  $r - p$  will be induced once we condition on  $\sum p_i = 1$ .

Regarding the choice of  $\delta^2$ , if we thought that the expert placed her chips to minimise  $\|r - p\|$  we could argue that an individual rounding error  $\varepsilon_i = r_i - p_i$  cannot be larger in magnitude than  $(k - 1)/(kn)$ . Such a rounding error could be achieved, for example, if  $\varepsilon_i = (k - 1)/(kn)$  and  $\varepsilon_j = -1/(kn)$  for  $j \neq i$ . In this case, the maximum error cannot be reduced: moving one chip from bin  $i$  to any other

bin  $j$  results in  $\varepsilon_i = -1/(kn)$  and  $\varepsilon_j = (k-1)/(kn)$ , leaving  $\|r-p\|$  unchanged. Alternatively, suppose we have  $\varepsilon_i = (k-1)/(kn) + \phi$ , with  $\phi > 0$ . Then at least one error,  $\varepsilon_j$  say, must be less than  $-1/(kn)$ , and so moving one chip from bin  $i$  to bin  $j$  reduces  $\|r-p\|$ .

Informally, it seems reasonable to suppose that a rounding error is likely to be in the interval  $(-1/n, 1/n)$ , but we choose not to rule out the possibility of larger errors; an expert might allocate one chip ‘too many’ (or ‘too few’) to a bin, while preferring not to put the chip elsewhere. We set  $\delta$  to be  $1/(2n)$ , and judge  $r$  to be independent of  $\alpha$  and  $\sigma^2$  given  $p$ .

Combining  $\pi(p|\alpha, \sigma^2 \sum p_i = 1)$  and  $\pi(r|p, \alpha, \sigma^2)$  we then have  $p|r, \alpha, \sigma^2 \sum p_i = 1 \sim N(H^{**}, A^{**})$ , with

$$A^{**} = (A^{*-1}/\sigma^2 + I/\delta^2)^{-1}, \quad H^{**} = A^{**}(A^{*-1}H^*/\sigma^2 + r/\delta^2).$$

Noting the indicator function in (1), we sample from  $\pi(p|r, \alpha, \sigma^2, \sum p_i = 1, p > 0)$  by sampling from  $N(H^{**}, A^{**})$  and rejecting any  $p$  with negative elements.

### 3.2.2 Sampling from $\pi(\alpha|p, \sigma^2)$ , $\pi(\sigma^2|p, \alpha)$ and $\pi(f|p, \alpha, \sigma^2)$ .

Conditional on  $p$ , we have data in the form considered by O&O, and so can follow their procedures for sampling from these three conditional distributions. Details are given in Appendix A.

## 4 Example 1

We consider a synthetic example, to investigate uncertainty about the expert’s density function resulting from different choices of the number of chips and bins used in the elicitation. We suppose that the expert’s true distribution is a  $Beta(9, 6)$  distribution. We consider combinations of 10 chips and 20 chips with 5 bins and 10 bins, equally spaced between 0 and 1. The allocation of chips to bins is shown in Table 1.

We suppose that the expert has stated that her distribution is unimodal, and so multimodal densities are rejected in the sampling process. Adjacent bins with 0 chips allocated are merged in the analysis. For example, in the 10 chips and 10 bins case, we consider a single rounding error, judged unlikely to be more than  $1/10$ , over the interval  $[0, 0.4]$ .

In Figures 2 and 3, we show the facilitator’s pointwise 95% intervals for the expert’s density and distribution functions (based on a sample of 1000 functions).

Bin	0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 - 0.4	0.4 - 0.5	0.5 - 0.6	0.6 - 0.7	0.7 - 0.8	0.8 - 0.9	0.9 - 1.0
10 chips, 10 bins	0	0	0	0	0.2	0.3	0.3	0.2	0	0
20 chips, 10 bins	0	0	0	0.05	0.15	0.25	0.3	0.2	0.05	0
10 chips, 5 bins	0		0.1		0.4		0.5		0	
20 chips, 5 bins	0		0.05		0.45		0.45		0.05	

Table 1: Implied probabilities given the four combinations of chips and bins. True probabilities are generated from a Beta(9,6) distribution.

For comparison, we also show what the pointwise 95% intervals would be were the expert to state her true probability for each bin. With 5 bins and noise-free probabilities, the facilitator’s uncertainty about the expert’s cdf is fairly small, so that a combination of a small number of bins and large number of chips can result in fairly little uncertainty about the expert’s density. This is to be expected: the Gaussian process model for  $f$  implies that the cdf is also a smooth function, so observing a small number of (suitably spaced) points on the cdf should reduce uncertainty substantially.

We observe less benefit in increasing the number of bins without increasing the number of chips. Some uncertainty has been reduced in the lower tail, but there is now potential for larger rounding errors over the same regions of the parameter space. For example, instead of considering one rounding error in the interval  $[0.4, 0.6)$  for the 5 bin case, we now consider rounding errors in each of the intervals  $[0.4, 0.5)$  and  $[0.5, 0.6)$  in the 10 bin case, where prior judgements of the likely magnitude of any single rounding error have not changed.

In Figure 4, we show the facilitator’s posterior distribution for each error ( $p_i - r_i$ ) in the cases of 5 bins and either 10 or 20 chips. The true errors are marked as vertical lines. This is to inspect the behaviour of the model, and check that posterior uncertainty about the expert’s distribution is based on plausible posterior distributions for the rounding errors.

Recall the prior judgement that rounding errors are likely to lie in the interval  $(-1/n, 1/n)$  (with truncation occurring for bins with no chips). The posterior distributions suggest errors well within these limits in most cases, with the exception of the bin  $[0.6, 0.8)$  in the case of 10 chips. Here, there is some possibility of a larger negative error. As the stated probability was 0 in the adjacent bin  $[0.8, 1)$ , the posterior distribution seems appropriate here.

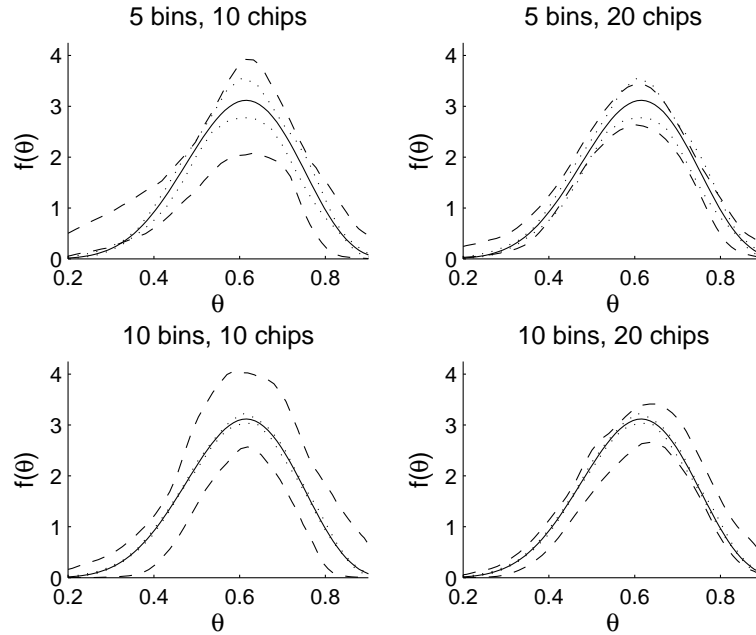


Figure 2: Pointwise 95% intervals for the expert's density function (dashed lines), and the Beta(9,6) density function (solid lines). Dotted lines show pointwise 95% intervals for  $f(\theta)$  given the precise probabilities  $p$ .

## 5 Example 2: CHART

We now consider an example reported in Parmar et al. (1994) and Spiegelhalter et al. (2004). This involves beliefs about the efficacy of a radiotherapy technique known as continuous hyperfractionated accelerated radiotherapy (CHART) in comparison with conventional radiotherapy, for lung cancer patients. The unknown parameter  $\theta$  of interest is the ratio of the hazard under CHART to the hazard under conventional radiotherapy.

The experts were asked to express their beliefs about the benefit from CHART, quantified as the percentage improvement in survival after two years. Specifically, they were each asked to distribute 100 points between 9 bins. The experts distributed their points in multiples of 5 or 10, and so for illustration, we suppose that each expert chose to allocate either 10 or 20 chips, with the choice of 10 or 20 reflecting the precision each expert was willing to consider. Hence we interpret the data as in Table 2

Noting the allocation of chips by experts six and seven, we suppose that these two experts were certain that CHART could not be worse than standard treatment, and so we treat the allocation of zero chips to the first three bins as a noise-free

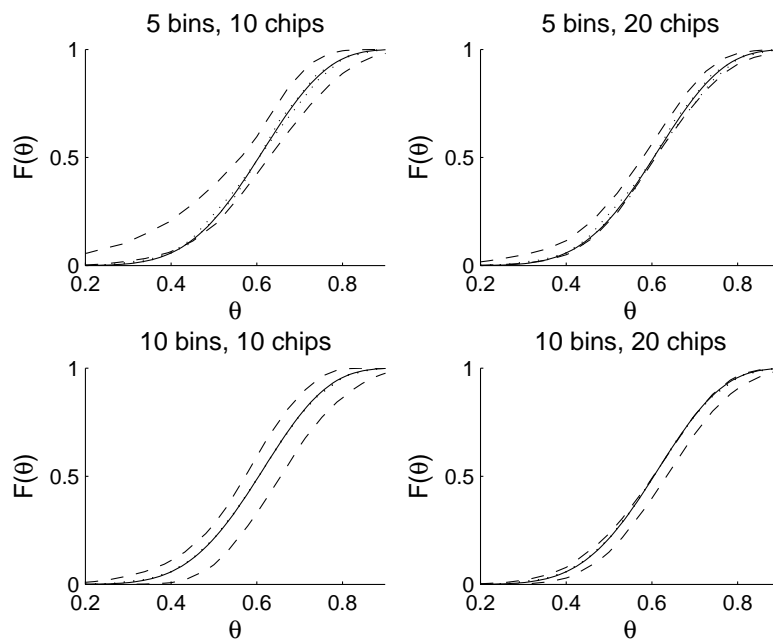


Figure 3: Pointwise 95% intervals for the expert's distribution function (dashed lines), and the Beta(9,6) distribution function (solid lines). Dotted lines show pointwise 95% intervals for  $f(\theta)$  given the precise probabilities  $p$  (these intervals are very small in the 10 bins case).

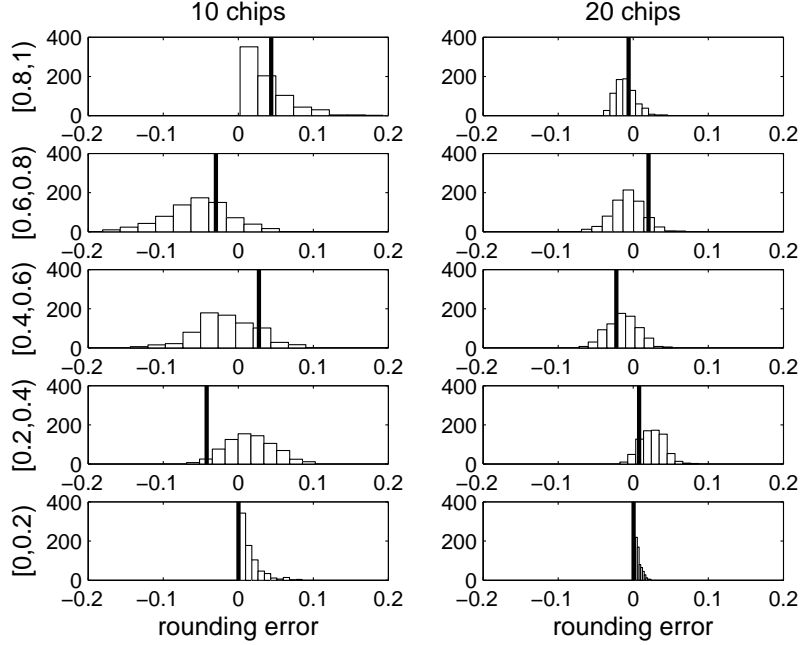


Figure 4: Posterior distribution for the rounding error ( $p_i - r_i$ ) in each bin, for 5 bins and 10 chips (1st column) and 20 chips (2nd column). The vertical line in each plot shows the true rounding error.

observation  $P(\theta < 0) = 0$ . Following Spiegelhalter et al. (2004), we convert the percentage improvements to hazard ratios assuming a 15% baseline survival.

Spiegelhalter et al. (2004) used linear pooling to obtain a consensus distribution  $f$  from the 11 elicited distributions, by fitting a lognormal distribution to the mean number of chips in each bin. Here, we consider the facilitator's uncertainty about this consensus distribution, allowing for uncertainty in the individual expert distributions. Defining  $r_{(i)}$  to be the allocation of chips for expert  $i$ , and  $R = (r_{(1)}, \dots, r_{(11)})$ , we simulate from  $\pi\{f|R\}$  as follows:

1. Fit the nonparametric model to expert  $i$ 's judgements, for  $i = 1, \dots, 11$ .
2. Generate a random density function  $f_{i,j}$  from  $\pi\{f_i|r_{(i)}\}$
3. Obtain a random draw from  $\pi\{f(\theta)|R\}$  as

$$\frac{1}{11} \sum_{i=1}^{11} f_{i,j}(\theta).$$

4. Repeat steps 2 and 3 a large number of times to obtain pointwise estimates and intervals for  $f(\theta)$ .

Expert	Total no. of chips	CHART worse than standard by %			CHART better than standard by %					
		10-15	5-10	0-5	0-5	5-10	10-15	15-20	20-25	25+
1	20	0	2	2	5	5	4	2	0	0
2	20	0	2	7	9	2	0	0	0	0
3	10	0	1	4	4	1	0	0	0	0
4	10	0	0	0	1	1	3	3	1	1
5	10	0	0	0	1	6	2	1	0	0
6	10	0	0	0	3	5	2	0	0	0
7	10	0	0	0	6	3	1	0	0	0
8	10	0	0	0	0	1	4	4	1	0
9	10	0	0	0	0	1	6	2	1	0
10	10	0	0	0	0	2	4	4	0	0
11	10	0	0	0	0	0	3	5	2	0

Table 2: Assumed allocation of chips to bins

The facilitator’s pointwise median and 95% intervals for the consensus distribution are shown in figure 5. For comparison, we also show the lognormal fit used by Spiegelhalter et al. (2004), together with the histogram based on the mean number of chips in each bin. The pointwise median appears to fit the histogram better than the lognormal distribution, although the difference is fairly minor. Here, uncertainty about  $f(\theta)$  is fairly small.

## 6 Discussion

In this paper we have coupled the nonparametric approach to elicitation of O &O with the roulette elicitation method. This allows us to quantify probabilistically uncertainty about the expert’s density, given her roulette judgements. The method is computationally intensive, and so would not be suitable for providing feedback to the expert in real time. However, once she has settled on her allocation of chips to bins, the method can be used to produce a sample of density functions for the analysis problem at hand. This will help test robustness to the choice of prior, though only as far as issues of imprecision in her assessments are concerned; one may still wish to investigate robustness to different prior beliefs.

In our synthetic example, we found that the combination of a small number of

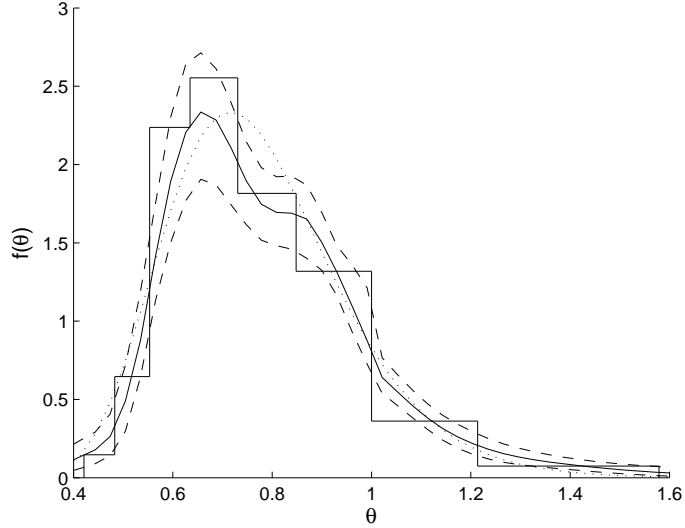


Figure 5: The facilitator’s pointwise median (solid line) and 95% intervals (dashed lines) for the consensus distribution. A lognormal fit (dotted line) and a histogram are shown for comparison.

bins with a large number of chips can result in fairly little uncertainty about the expert’s distribution. However, if an expert feels she cannot provide probabilities very precisely, then a smaller number of chips may be preferable, so that her imprecision is properly recognised. In such cases, it will be more important to investigate uncertainty about her distribution using the nonparametric approach. Although our method is complex to implement, given the difficulty of making probability assessments in general, we think methods that transfer some of the effort from the expert to the facilitator are highly desirable.

## Appendix A

- Sampling from  $\pi(\alpha|p, \sigma^2)$

In the case of a  $N(m, v)$  density function for  $g(\theta)$  and the modelling choices given in 2, O&O show that the posterior density of  $\alpha$  conditional on  $p$  and  $\sigma^2$  is given by

$$\begin{aligned} \pi(\alpha|p, \sigma^2) &= \pi(m, v, b^*|p, \sigma^2) \\ &\propto v^{-1}|A|^{-\frac{1}{2}} \frac{1}{b^*} \exp \left\{ -\frac{1}{2}(\log b^*)^2 - \frac{1}{2\sigma^2}(p - H)^T A^{-1}(p - H) \right\}, \quad (4) \end{aligned}$$

We use the Metropolis-Hastings algorithm to sample from (4), as described in O&O.

- Sampling from  $\pi(\sigma^2|p, \alpha)$   
If  $\sigma^{-2} \sim \text{Gamma}(d, a)$ , then

$$\sigma^{-2}|p, \alpha \sim \text{Gamma}\{d + n/2, a + 0.5(p - H)^T A^{-1}(p - H)\},$$

and we can sample from this conditional directly.

- Sampling from  $\pi(f|p, \alpha, \sigma^2)$

We sample a finite set of points  $d = \{f(\theta_1), \dots, f(\theta_s)\}$  on the expert's density function, which is straightforward to do as the distribution of  $d, p, \alpha$  (and therefore  $d|p, \alpha$ ) is multivariate normal. Note that if  $g(\theta|\alpha)$  is the normal density function with mean  $m$  and variance  $v$ , then the covariance between  $f(\theta)$  and  $p_i$  is

$$\begin{aligned} \text{cov}\{p_i, f(\theta)|\alpha, \sigma^2\} &= \sigma^2 g(\theta) \int_{-\infty}^x c(\theta, \theta') g(\theta') d\theta' \\ &= \sigma^2 g(\theta) \left( \frac{b^*}{1 + b^*} \right)^{\frac{1}{2}} \exp \left\{ -\frac{(\theta - m)^2}{2v(b^* + 1)} \right\} \\ &\times \left[ \Phi \left\{ \left( x_i - \frac{\theta + mb^*}{b^* + 1} \right) \sqrt{\left( \frac{1 + b^*}{vb^*} \right)} \right\} - \Phi \left\{ \left( x_{i-1} - \frac{\theta + mb^*}{b^* + 1} \right) \sqrt{\left( \frac{1 + b^*}{vb^*} \right)} \right\} \right], \end{aligned}$$

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

For suitably chosen  $\theta_1, \dots, \theta_s$  (arranged in increasing order) we will have  $\text{Var}\{f(\theta)|d, p, \alpha\}$  negligibly small for any  $\theta \in [\theta_1, \theta_s]$ , and so we can use  $E\{f|d, p, \alpha\}$  as an approximate draw from the distribution of  $f|p, \alpha$ . (For the choice of  $\theta_1, \dots, \theta_s$ , typically no more than 10 evenly spaced values over the region of interest are required, as we are only trying to learn a smooth function of a single input variable). It is at this stage that we apply the constraint  $f \geq 0$  by discarding any sample  $d$  for which  $E\{f(\theta)|d, p, \alpha\}$  is negative at any  $\theta$  of interest.

## 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*, edited by Kahneman, D., Slovic, P. and Tversky, A., Cambridge: Cambridge University Press.
- Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior, *Journal of Statistical Planning and Inference*, **25**: 303–328.
- Bowman, A. W. and Crawford, E. (2008). *R package rpanel: simple control panels (version 1.0-5)*, University of Glasgow, UK.
- Daneshkhah, A. and Oakley, J. E. (2010). Eliciting multivariate probability distributions, in *Rethinking Risk Measurement and Reporting: Volume I*, edited by Böcker, K., Risk Books, London.
- Garthwaite, P. H., Jenkinson, D. J., Rakow, T. and Wang, D. D. (2007). Comparison of fixed and variable interval methods for eliciting subjective probability distributions, Tech. rep., University of New South Wales.
- Garthwaite, P. H., Kadane, J. B. and O’Hagan, A. (2005). Statistical methods for eliciting probability distributions, *J. Am. Statist. Assoc.*, **100**: 680–701.
- Good, I. J. (1980). Some history of the hierarchical Bayesian methodology (with discussion), in *Bayesian Statistics*, edited by Bernardo, J. M., Groot, M. H. D., Lindley, D. V. and Smith, A. F. M., pp. 481–519, Valencia: University Press.
- Gore, S. M. (1987). Biostatistics and the Medical Research council, *Medical Research Council News*.
- Gosling, J. P., Oakley, J. E. and O’Hagan, A. (2007). Nonparametric elicitation for heavy-tailed prior distributions, *Bayesian Analysis*, **2**: 693–718.
- Hughes, M. D. (1991). Practical reporting of bayesian analyses of clinical trials, *Drug information journal*, **25**: 381–93.
- Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A. and m. Feldman, B. (2010). A valid and reliable belief elicitation method for bayesian priors, *Journal of Clinical Epidemiology*, **63**: 370–383.

- Moala, F. M. and O'Hagan, A. (2010). Elicitation of multivariate prior distributions: a nonparametric bayesian approach, *J. Statist. Plan. and Infer.*, **140**: 1635–1655.
- Murphy, A. H. and Winkler, R. L. (1974). Credible interval temperature forecasting: some experimental results, *Monthly Weather Review*, **102**: 784–794.
- Oakley, J. E. (2010). Eliciting univariate probability distributions, in *Rethinking Risk Measurement and Reporting: Volume I*, edited by Böcker, K., Risk Books, London.
- Oakley, J. E. and O'Hagan, A. (2007). Uncertainty in prior elicitation: a nonparametric approach, *Biometrika*, **94**: 427–441.
- Oakley, J. E. and O'Hagan, A. (2010). *SHELF: the Sheffield Elicitation Framework (version 2.0)*, School of Mathematics and Statistics, University of Sheffield, <http://tonyohagan.co.uk/shelf>.
- 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 (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester: Wiley.
- 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.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall.