ELICITING AND COMBINING EXPERT OPINION - AN OVERVIEW AND COMPARISON OF METHODS

by

MUTSA CAROLE CHINYAMAKOBVU

A thesis submitted in partial fulfillment of the requirements for the degree

Master of Science

in

Mathematical Statistics

Department of Statistics

Science Faculty

Rhodes University

Supervisor: Dr. I. Garisch

Grahamstown

January 2015

Declaration

I hereby declare that this thesis is my original work and where other people's work has been used due reference has been given.

Date: ....

Student's signature: ....

Supervisor's signature: ....

Abstract

Decision makers have long relied on experts to inform their decision making. Expert judgment analysis is a way to elicit and combine the opinions of a group of experts to facilitate decision making. The use of expert judgment is most appropriate when there is a lack of data for obtaining reasonable statistical results.

The experts are asked for advice by one or more decision makers who face a specific real decision problem. The decision makers are outside the group of experts and are jointly responsible and accountable for the decision and committed to finding solutions that everyone can live with. The emphasis is on the decision makers learning from the experts.

The focus of this thesis is an overview and comparison of the various elicitation and combination methods available. These include the traditional committee method, the Delphi method, the paired comparisons method, the negative exponential model, Cooke's classical model, the histogram technique, using the Dirichlet distribution in the case of a set of uncertain proportions which must sum to one, and the employment of overfitting. The supra Bayes approach, the determination of weights for the experts, and combining the opinions of experts where each opinion is associated with a confidence level that represents the expert's conviction of his own judgment are also considered.

Keywords: Expert judgment analysis, traditional committee method, Delphi method, paired comparisons method, negative exponential model, Cooke's classical model, Dirichlet distribution, Supra Bayes approach

Acknowledgements

I would like to thank my parents for their support throughout this work. I would like to also thank the staff and fellow students in the Department of Statistics at Rhodes University for their boundless help and support. I am especially grateful for my supervisor, Dr. Isabelle Garisch, for her guidance and generous contribution of time and expertise towards this thesis.

Special mention goes to Dr. Lizanne Raubenheimer from Rhodes University, and Prof. Thomas Mazzuchi from George Washington University, for always being willing to lend a hand, some knowledge, some extra material, and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and significance of study

Decision analysis is an increasingly popular field of study of the 21^{st} century. The rapidly advancing world of technology is generating large amounts of data on various platforms through logs of cellphone use, credit cards and internet history, to name a few. Google can predict what you are searching for from just the first few letters typed into the search-bar and from logs of the websites you regularly visit. According to Mayer-Schonberger and Cukier (2013), it can even collate various users' searches to predict where a flu will spread to. This growing trend of massive amounts of seemingly stagnant data informing decisions by facilitating predictions beyond real time has been awarded the term "Big Data". Data scientists who have an aptitude for statistics and databases are considered the scarce and invaluable experts of "Big Data".

On the other end of the scale, there is growing interest in fields where there are a *lack* of data to produce reasonable results through normal statistical methods. These fields include climate change, accident consequence management for nuclear power plants and other critical infrastructures, aircraft wiring risk assessment, cyber security, maintenance optimisation, to mention but a few. Expert Judgment Analysis is a technique that draws opinions from experts in the field of interest about the unknown variables. Eggstaff et al. (2014) explain that these opinions are weighted and combined to represent the uncertainty or risk in decision analysis by an external decision maker who is committed to determine the best possible solution from the experts. Various methods of eliciting and combining expert opinion have been developed over the years and Cooke's classical model has been a popular approach for over 20 years. Since fields that lack data are prone to high levels of risk, expert judgment analysis could be considered "Big Data" 's coy mistress. Cooke and Goosens (2010) explain that the experts may have valuable but uncertain (due to the lack of data) knowledge on the parameters of interest, which can be specified to some subjective degree of belief. It is then up to the decision maker to ultimately produce "optimally defensible" choices of these parameters.

1.2 Objectives of study

The objectives of this thesis are:

- to give an overview and comparison of various methods for elicitation and combination of expert opinion currently available. These include the Traditional Committee method, the Delphi method, the Paired Comparisons method, the Cooke method and using the Dirichlet distribution;

- to demonstrate the most popular method of elicitation, namely, Cooke's "Classical model".

1.3 Organisation of study

Chapter 2 discusses various methods of eliciting and combining expert opinion. As mentioned in the abstract, these include the traditional committee method, the Delphi method, the paired comparison method, the Cooke method and using the Dirichlet distribution. Some models and techniques that employ the approaches of some of these methods are also discussed, namely the negative exponential life model, the classical model, the histogram technique and the supra Bayes approach. Chapter 2 ends off with a section on some instances of the practical application of expert judgment analysis. Chapter 3 presents our own application of some of the methods discussed in Chapter 2. The results will be compared and discussed as per the objectives of this study. Finally, Chapter 4 concludes the findings of this study and suggest areas of further research.

Chapter 2

Expert Judgment Analysis

2.1 General Overview

Simply put by Ryan et al. (2012), expert judgment analysis has been introduced as a way to elicit, codify and pool the opinions of a select group of specialists in the face of uncertainty. According to Cooke and Goosens (2010), the group of specialists may possess valuable knowledge about models and parameters in the subject matter, and they can draw from their vast expertise in their particular field of interest to assess unknown quantities. The knowledge they can offer is not certain but the specialists can specify "degrees of belief" that can be quantified and aggregated to give optimal choices of parameters and models. Note that since the experts themselves are not certain, they typically will not agree (Cooke and Goosens (2008)), so expert agreement should not be a goal of this analysis. They also emphasise that for expert judgment analysis to be applicable to a problem, relevant scientific expertise, along with the relevant theories and measurements, must already exist. The specific quantities of interest are the only variables that can be unmeasurable in practice. Cooke and Goosens (2010) explain that techniques can be applied to give quantitative assessments of uncertainty, where the assessments are given as subjective probability distributions from which nominal values of the parameters of interest can be derived from. One could also apply techniques to give qualitative and comparative assessments where the resulting alternatives are ranked. The choice of technique is influenced by the scrutability of the data and processing tools, fairness among the experts, neutrality of the methods against bias, and performance control of the various techniques.

Cooke and Goosens (2008) determined that expert judgment can aim to produce, in order of strength of analysis type; census, political consensus or rational consensus. If the aim is to produce a "census", they explain that the distribution of views within the expert community are simply surveyed, and differential weighting can be considered in order to ensure that the expert community is represented accordingly. If the aim is to produce "political consensus", the experts are allocated weights reflective of the interests or stake-holders they will be representing. If each stakeholder is represented by the same number of experts then equal weighting can be applied. They describe “rational consensus” as a group decision process where the experts, without knowing the final outcome, agree on how to represent and combine their uncertainties.

Cooke and Goosens (2010) point out that a distinctive feature of the expert judgment analysis methods developed by Delft University of Technology is control assessments. These are assessments of uncertain quantities that resemble the variables of interest closely, where the actual values are known after the experiment. Control assessments can be applied to both expert- and combinations of expert assessments, and they provide the basis of performance criteria that diverse stakeholders will look at when deciding whether to buy into the process.

EMSE280 (2011) states that the use of expert judgment is most appropriate when there is a lack of quantitative data, or reliable good quality data, for obtaining reasonable statistical results. In some instances the data may be available but it may not have been collected with risk analysis in mind, necessitating the use of a specialised method of analysis. Other variables to consider when executing this process are travel, prior training undergone by experts on subjective probability assessments and the level of documentation available on the subject on interest. In addition, Walker et al. (2004) notes that financial constraints on the project, availability of the experts and any institutional affiliations they may have must also be considered before elicitation. Software support is also key during the processing and write-up of the results so the facilitator would have to consider this prior to the elicitation. All these concerns will have to be considered by the facilitator. Walker et al. (2004) warns that the incentive for experts in elicitation should not be to perform as well as they can, as that will introduce bias and deviate from the required and informative truths.

2.2 Expert Elicitation

To be able to use performance-based elicitation, it is important to consider carefully how to select the experts and how many are necessary for this procedure. Aspinall (2010) notes that the client may want a range (with regards to field of expertise) of experts to achieve a panel as impartial as possible, or may restrict the panel to employees within the company to maintain confidentiality and ensure relevance. Cooke and Goosens (2010) write that previous studies in this area of research have used between 4 and 20 experts, and the process duration can be anything from a month to a year. Typically, in the elicitation phase of expert judgment analysis, one expert is used to conduct a dry run before finalising the elicitation questions. A meeting is then held with all the experts present to explain the study design and how the scoring and combining will work. The experts then undergo individual elicitation where they are asked to answer carefully administered seed questions. These seed questions are used to calibrate the performance of the experts and subsequently weight their assessments and opinions before combining. The above mentioned authors add that these seed variables can also be used to evaluate and validate the combined expert assessments as they enable “empirical control of any combination schemes”. They also emphasise the importance of ensuring that performance on seed variables is pre-judged relevant for performance on the variables of interest. So overall, seed variables aim to:

- evaluate experts’ subjective probability assessments to determine how they perform;
- facilitate combining expert distributions in a way that optimises performance; and
- gauge and ultimately justify the combination of expert assessments.

The process of expert elicitation has developed and evolved over the years and there are now various methods and models to consider when carrying out expert judgment analyses.

2.2.1 The traditional committee method

By Cooke and Goosens (2010)’s definition, the traditional committee is a behavioural aggregation method where the experts interact to achieve homogeneity of information on the variables of interest. Some behavioural approaches aim to get the experts to agree on a final probability density function for each variable, while in others, the experts assessments are combined by equal weighting to get the single probability density function per variable. Cooke and Goosens (2010) argue that the results produced prove to be inferior to results from elicitation methods that use a mathematical approach such as the Cooke method discussed in section 2.2.4.

Oblivious to its flaws, many sectors still use the traditional committee method as it is the most obvious approach before careful thought. A wide range of opinions from various experts are gathered over a slow deliberative process. Each participant is allowed one vote, resulting in the experts all assuming equal weights. Aspinall (2010) highlights that the downfall with this method is the inherent assumption that each expert is equally proficient, equally informed and unbiased. In the analysis being reviewed by Walker et al. (2004), it was also noted that experts were being considered as though they were interchangeable, sharing a common variance and a common degree of correlation, when in reality, pairs of experts are likely to differ.

2.2.2 The Delphi method

Eggstaff et al. (2014) describe the Delphi method as a behavioural technique that is prone to psychological bias that may affect the validity of the results. The Delphi method requires each expert to give a formal statement regarding their position on the matter at hand. Cooke and Goosens (2010) explain that the statement must specify guesses of the values of the unknown quantities as single point estimates, which makes this method quick and easy to execute. The statements are then circulated several times and experts are allowed the opportunity to adjust their opinions. After observation, the guesses are then compared with the observed values. Due to the several rounds of circulation, de Franca Doria et al. (2009) warn that the decision maker could end up with a long list of issues that can be time-consuming and difficult to collate and reach a consensus. de Franca Doria et al. (2009) handled this problem in their study by redefining the definitions in each round then stopping the interactions when their pre-determined level of 80% consensus was reached.

During the circulation of the statements, as can be expected in group dynamics, Cooke and Goosens (2010) and Aspinall (2010) point out that adjustments will tend to follow the direction of the supposed "leading" expert of the group as opposed to being steered by the strongest argument. Aspinall (2010)'s practical application of Cooke's method described in section 2.3 shows that neither self confidence nor scientific prestige and reputation can be considered good predictors of expert performance.

Cooke and Goosens (2010) note that one of the reasons this method is no longer common is that the observed values and guessed estimates do not have any base scale that would warrant useful comparison and interpretation. Specifying the guesses as point estimates unfortunately means that no indication of uncertainty is provided with the estimate. Furthermore, the methods used to process and combine the opinions inherently and incorrectly assume that the measurements are normal physical measurements.

2.2.3 The paired comparisons method

In the paired comparison method, Cooke and Goosens (2010) explain that alternatives are compared and ranked pairwise based on a criterion such as preference, feasibility, etc, making sure that each item involved is compared with all the other items in turn. This unfortunately opens the method up to redundancy. Another disadvantage of this method is that there is no assessment of uncertainty, as with the Delphi and Traditional Committee methods. However, depending on the availability of necessary observed values and the method chosen for ranking, these authors propose that the data can actually be further reduced to an interval or a ratio scale, thus making this method a popular choice for expert elicitation.

The Negative Exponential Life model

According to Mazzuchi et al. (2008), a popular model based on the paired comparison method for expert judgment analysis is the Negative Exponential Life (NEL) model, where n components or environments are compared pairwise then ranked. One can then analyse whether each expert is specifying a true preference structure in his/her answers or merely assigning answers randomly. This can be done by analysing the count of circular triads in his/her comparisons. A circular triad happens when one suggests that option 1 is better than option 2, and option 2 is better than option 3, and then option 3 is better than option 1. When a vast number of events are being compared, such circular triads may occur.

When testing whether all the experts' agreements are due to chance, the coefficient of agreement for comparing n items among e experts is defined as

$$u = \frac{2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\binom{N(i,j)}{2}}{\binom{e}{2}\binom{n}{2}} - 1, \tag{2.1}$$

where $N(i,j)$ represents the number of times an expert ranked environment E_i as more 'severe' than environment E_j. Tabulated distributions of $\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\binom{N(i,j)}{2}$ for small values of n and e under the null hypothesis that all the agreements of the experts are attributed to chance have been developed, and the hypothesis concerning u can be tested using these distributions. As an illustration, suppose we are comparing $n=4$ items among $e=2$ experts. Thus

$$\begin{aligned} u &= \frac{2\sum_{i=1}^{4}\sum_{j=1,j\neq i}^{4}\binom{N(i,j)}{2}}{\binom{2}{2}\binom{4}{2}} - 1 \\ &= \frac{2}{\binom{2}{2}\binom{4}{2}}\left[\binom{N(1,2)}{2}+\binom{N(1,3)}{2}+\binom{N(1,4)}{2}+\binom{N(2,1)}{2}\right. \\ &\quad +\binom{N(2,3)}{2}+\binom{N(2,4)}{2}+\binom{N(3,1)}{2}+\binom{N(3,2)}{2}+\binom{N(3,4)}{2} \\ &\quad \left.+\binom{N(4,1)}{2}+\binom{N(4,2)}{2}+\binom{N(4,3)}{2}\right] - 1. \end{aligned}$$

Suppose expert 1's rankings of (i,j) are as follows: $1>2$, $1>3$, $4>1$, $2<1$, $2>3$, $2<4$, $3<1$, $3<2$, $3<4$, $4>1$, $4>2$, $4>3$; thus, $4>1>2>3$.

Furthermore, suppose that expert 2's rankings of (i,j) are as follows: $1<2$, $1<3$, $1<4$, $2>1$,

$2 < 3$, $2 < 4$, $3 > 1$, $3 > 2$, $3 > 4$, $4 > 1$, $4 > 2$, $4 < 3$; thus, $3 > 4 > 2 > 1$.

According to these rankings, the values of $N(i, j)$ are given in the following table:

		i			
		1	2	3	4
j	1	-	2	2	0
	2	2	-	2	0
	3	2	2	-	2
	4	4	4	2	-

From this table, it follows that:

$$\begin{aligned} u &= \frac{2}{\binom{2}{2}\binom{4}{2}}\left[\binom{2}{2}+\binom{2}{2}+\binom{2}{2}+\binom{2}{2}+\binom{2}{2}\right. \\ &\quad \left.+\binom{2}{2}+\binom{2}{2}+\binom{4}{2}+\binom{4}{2}+\binom{2}{2}\right]-1 \\ &= 5.6667\,. \end{aligned}$$

The minimum value of u indicating “no agreement” would be given by

$$\begin{aligned} u_{min} &= \frac{2}{6}(12)-1 \\ &= 3, \end{aligned}$$

and the maximum value of u indicating “total agreement” would be given by

$$\begin{aligned} u_{max} &= \frac{2}{6}(36)-1 \\ &= 11. \end{aligned}$$

In general, “no agreement” can be respresented by:

$$\begin{aligned} u_{min} &= \frac{2n(n-1)}{\begin{pmatrix} e \\ 2 \end{pmatrix}\begin{pmatrix} n \\ 2 \end{pmatrix}} - 1 \\ &= \frac{2n(n-1)}{\begin{pmatrix} e \\ 2 \end{pmatrix}\left(\frac{n(n-1)}{2}\right)} - 1 \\ &= \frac{4}{\begin{pmatrix} e \\ 2 \end{pmatrix}} - 1, \end{aligned} \tag{2.2}$$

which is independent of the number of itens n. The following graph shows the minimum coefficient of agreement u_{min} (indicating “no agreement”), versus the number of experts e, where $e = 1, 2, ..., 20$.

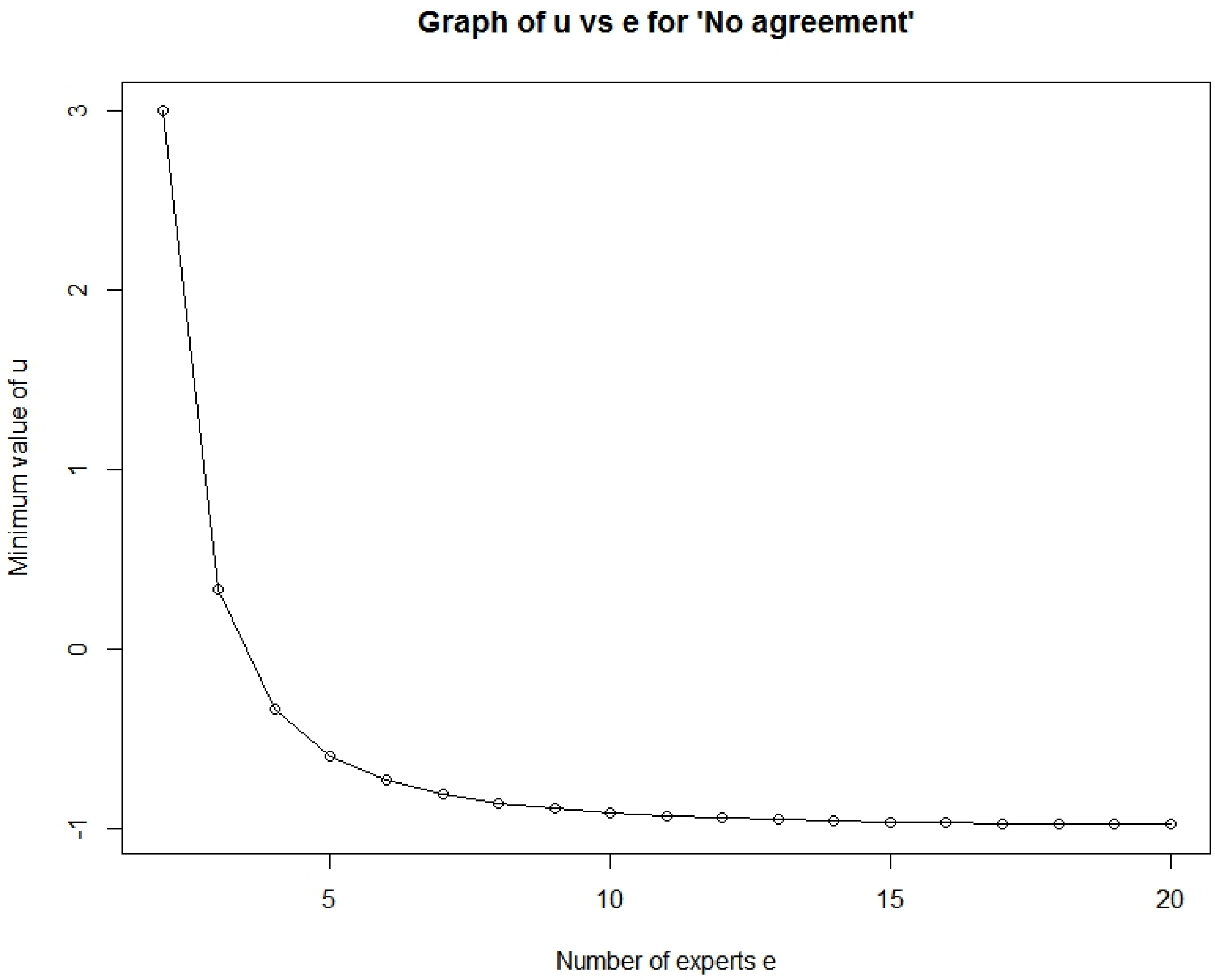


Figure 2.1: No agreement

Figure 2.1 shows that the minumum value u_{min} decreases as e increases. For $e > 10$, $u_{min} \to -1$.

"Total agreement" can, in general, be represented by:

$$\begin{aligned} u_{max} &= \frac{2\left(\frac{n(n-1)}{2}\binom{n}{2}\right)}{\binom{e}{2}\binom{n}{2}} - 1 \\ &= \frac{n(n-1)}{\binom{e}{2}} - 1. \end{aligned} \tag{2.3}$$

Note that while u_{min} is independent of n, this is not the case for u_{max}. For a range of 1 to 20 experts, “total agreement” as represented by equation 2.3 can be graphed as shown by figure 2.2.

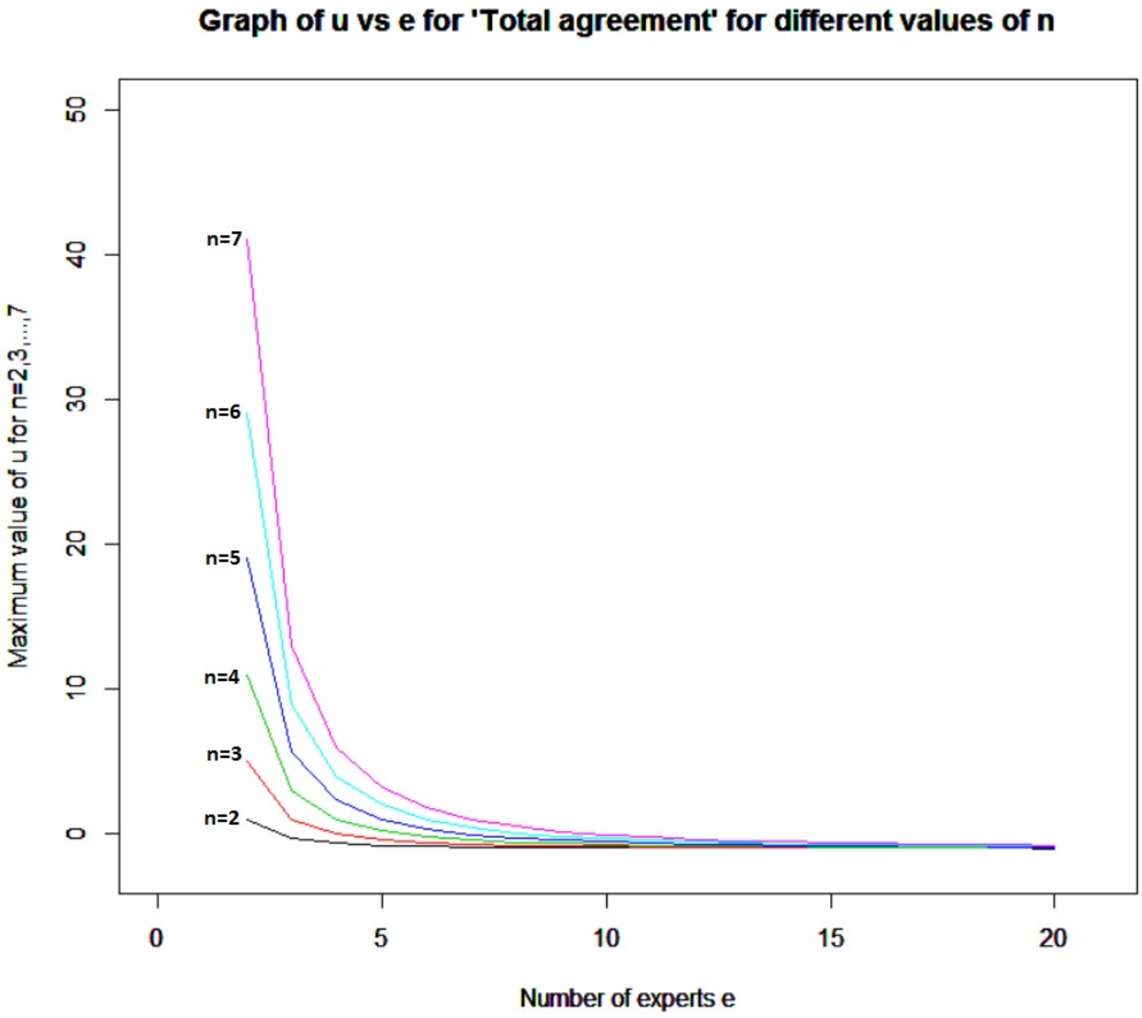


Figure 2.2: Total agreement

As shown in figure 2.2, as long as we have a few (less than 5) experts, u_{max} increases as n increases.

Mazzuchi et al. (2008) explains that the coefficient of concordance can be used as a measure of agreement among the experts, and it is defined as

$$w = \frac{S}{\frac{1}{12}e^2(n^3 - n)}, \tag{2.4}$$

or

$$w' = \frac{S}{\frac{1}{12}en(n+1)} \quad \text{for } n > 7, \tag{2.5}$$

where

$$S = \sum_{i=1}^{n} \left[\sum_{r=1}^{e} R(i,r) - \frac{\sum_{j=1}^{n}\sum_{r=1}^{e} R(j,r)}{n} \right]^2 , \tag{2.6}$$

and $R(j,r)$ is the rank of environment E_j obtained through expert r's responses, where $r = 1,...,e$. Equation 2.4 yields 1 when there is "total agreement" among the experts. Equation 2.5 was developed by Siegel (1956) and is said to have a chi-squared distribution with $n-1$ degrees of freedom. In order to have confidence in the estimates given by the experts, the one-tailed hypothesis that all the agreements can be attributed to chance for both w and w' should be rejected at a level of significance of 5%. For complete agreement, w attains the value 1.

To compare the two statistics w and w', consider the ratio

$$\begin{aligned} \frac{w}{w'} &= \frac{\left(\frac{S}{\frac{1}{12}e^2(n^3-n)}\right)}{\left(\frac{S}{\frac{1}{12}en(n+1)}\right)} \\ &= \frac{\left(\frac{1}{e(n^2-1)}\right)}{\left(\frac{1}{(n+1)}\right)} \\ &= \frac{1}{e(n-1)}. \end{aligned}$$

Let $k = \frac{1}{e(n-1)}$, thus $w = kw'$. For large values of e or n, $k \to 0$. The maximum value of k is 0.5 when $e = 2$ and $n = 2$. Therefore $0 < k < 0.5$ and $w > w'$.

Continuing with the example of comparing 4 items among 2 experts, S can be calculated from equation 2.6. First we determine the values of $\sum_{r=1}^{2} R(i,r)$ for $i = 1,...,4$:

$$\begin{aligned} \text{For } i = 1 &: \; R(1,1)+R(1,2) = 2+4 = 6; \\ \text{For } i = 2 &: \; R(2,1)+R(2,2) = 3+3 = 6; \\ \text{For } i = 3 &: \; R(3,1)+R(3,2) = 4+1 = 5; \\ \text{For } i = 4 &: \; R(4,1)+R(4,2) = 1+2 = 3. \end{aligned}$$

Then:

$$\begin{aligned}\frac{\sum_{j=1}^{4}\sum_{r=1}^{2}R(j,r)}{4} &= \frac{R(1,1)+R(1,2)+R(2,1)+R(2,2)+R(3,1)+R(3,2)+R(4,1)+R(4,2)}{4}\\ &= 5.\end{aligned}$$

From this it follows that:

$$\begin{aligned}S &= \sum_{i=1}^{n}\left[\sum_{r=1}^{e}R(i,r)-\frac{\sum_{j=1}^{n}\sum_{r=1}^{e}R(j,r)}{n}\right]^2\\ &= \left[(6-5)^2+(6-5)^2+(5-5)^2+(3-5)^2\right]\\ &= 6.\end{aligned}$$

Figure 2.3 is a graphical representation of the coefficient of concordance, w and w', for $n=(5,6,...,10)$ and $e=(2,3,...,10)$, using $S=6$:

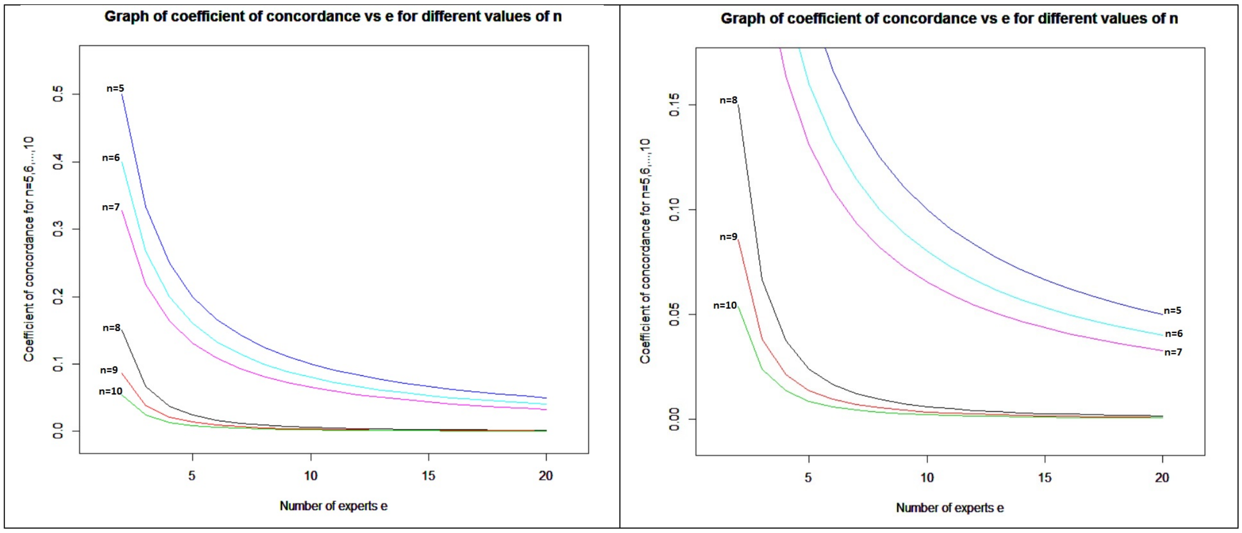


Figure 2.3: Coefficients of concordance, w and w' (alongside a zoomed-in version on the right)

As shown in figure 2.3, w and w' decrease as n increases. We also see that the graphs for $n=8,9\,\text{and}\,10$ converge at approximately $e=15$, and $w'\to 0$. This decrease is more evident for small values of e.

It is worth noting that, for our particular illustration, using w' when $n>7$ seems to be of little importance. In fact, figure 2.4 shows that using w when $n>7$ instead of w' produces higher (and thus more preferable) values for the coefficient of concordance.

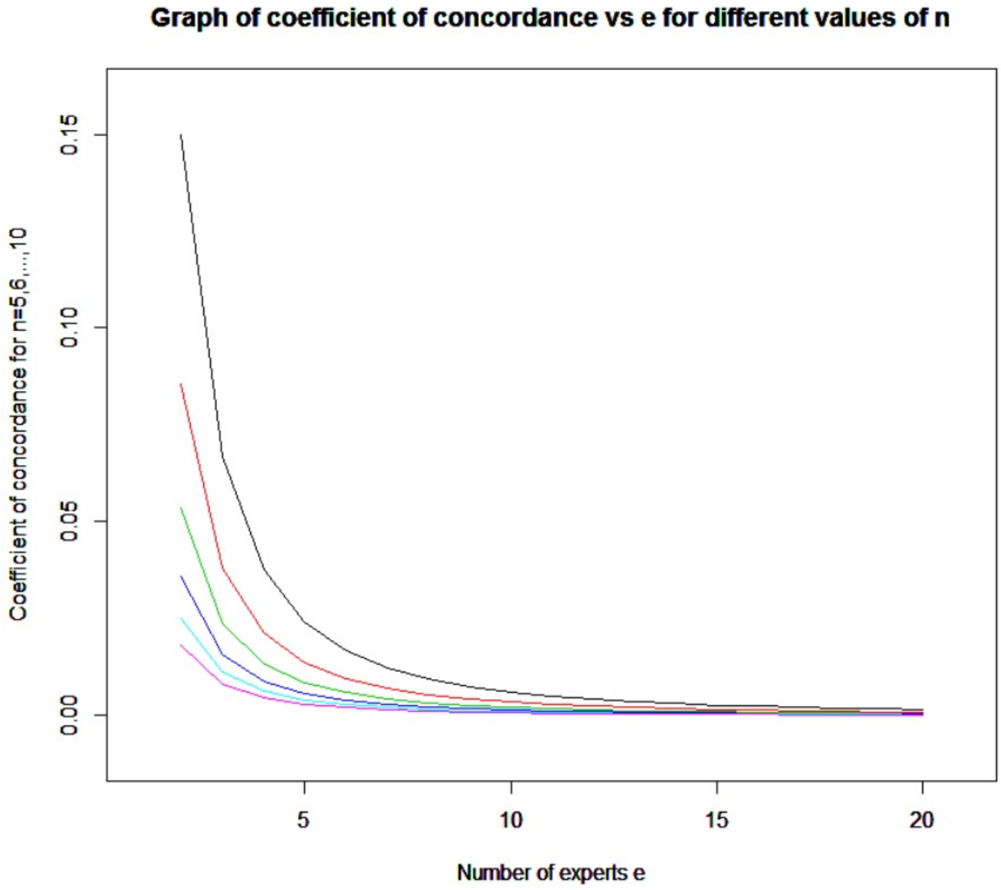


Figure 2.4: Coefficient of concordance, w, for $n = 5, 6, .., 10$.

The NEL model uses the idea that when comparing two environments E_i and E_j with respective rates of a quantity of interest h_i and h_j, the probability that E_i produces the quantity of interest before E_j is given by

$$r(i,j) = \frac{h_i}{h_i + h_j}. \tag{2.7}$$

Again, continuing with the example, if 1^{st}, 2^{nd}, 3^{rd} and 4^{th} are the rates h_i; and the ranked environments E_i for expert 1 are $4 > 1 > 2 > 3$, respectively, then equation 2.7 yields the following probabilities $r(i, j)$:

		i			
		1	2	3	4
j	1	-	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{1}{3}$
	2	$\frac{2}{5}$	-	$\frac{4}{7}$	$\frac{1}{4}$
	3	$\frac{1}{3}$	$\frac{3}{7}$	-	$\frac{1}{5}$
	4	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{4}{5}$	-

Similarly, using the same rates h_i as above; and letting the environments E_i for expert 2 be $3 > 4 > 2 > 1$, respectively, then equation 2.7 yields the following probabilities $r(i, j)$:

		i			
		1	2	3	4
j	1	-	$\frac{3}{7}$	$\frac{1}{5}$	$\frac{1}{3}$
	2	$\frac{4}{7}$	-	$\frac{1}{4}$	$\frac{2}{5}$
	3	$\frac{4}{5}$	$\frac{3}{4}$	-	$\frac{2}{3}$
	4	$\frac{2}{3}$	$\frac{3}{5}$	$\frac{1}{3}$	-

As shown above, $r(i, j) + r(j,i) = 1, \forall i, j, i \neq j$.

It is worth mentioning that the NEL model was derived from the Bradley-Terry model (Mazzuchi et al. (2008)).

2.2.4 The Cooke method

When faced with hard-to-assess risks, Aspinall (2010) writes that weighting the opinion of each expert based on their knowledge and ability to judge uncertainties relevant to the problem can produce a "rational consensus". Ryan et al. (2012) explain that the classical method of Cooke was developed towards this research effort of obtaining a "rational consensus", and it has been used extensively in the field of risk analysis. According to Cooke and Goosens (2010), mathematical aggregation methods are used to combine the opinions on each variable by applying analytical procedures that operate on the individual opinions. Such mathematical techniques are not prone to psychological bias as are behavioural techniques like the Delphi and the traditional committee methods. Aspinall (2010) chose the Cooke method for expert elicitation as it can bring to light disparities of opinion and loopholes in knowledge by tackling the problem one small section at a time. In reality, total consensus is said to be highly improbable and Cooke warns that trying to impose agreement will "promote confusion between consensus and certainty", as discussed in section 2.1. Instead of tailoring the decision-making process to reach an agreeable result, the goal should rather be to quantify the uncertainty among the experts and use that as the basis for the decision. This perspective allows for the consideration of the variation among the experts, and when data are sparse, unreliable or unobtainable, Aspinall (2010) considers the Cooke method to be most effective.

Cooke and Goosens (2010) give the following brief summary of the procedure and protocol for an expert judgment exercise:

Preparation for elicitation

1. Definition of case structure

   This is a document that provides the framework for the panel of experts specifying all issues to be taken into consideration during the exercise, including what is expected of the experts and the aim of the exercise. The document also gives some background information to the exercise, explains any physical phenomena and models the experts will be working with and states the conditions for the quantitative assessments.

2. Identification of target variables

   This step involves identifying and creating a list of all the model parameters and variables to be assessed by the experts. The most important parameters are then selected by means of some formal procedure, bearing in mind the lack of historical data that paves way for structured expert judgment.

3. Identification of query variables

   The target variables from the previous step may not be appropriate for elicitation so query variables are posed to the experts. During this step it is important to request observable quantities from the experts, and to structure the questions in a way that is consistent with how the experts represent relevant information.

4. Identification of performance variables

   These are handled similarly to query questions and supported with experimental evidence unknown to the panel of experts but known to the analyst. Performance variables are also referred to as “seed” or “domain” variables. These are the variables used to generate each expert’s weights.

5. Identification of experts

   Experts should be identified from fields relevant to the subject in question. These experts are sometimes referred to as “domain” experts. If they’re knowledgeable in subjective probability they referred to as normative experts.

6. Selection of experts

   From the list of possible experts for the panel created in the previous step, the largest number of experts possible are selected for elicitation. The selection criteria are:

   - reputation and experimental experience in the field of interest
   - number and quality of publications in the field of interest
   - diversity and background
   - familiarity with uncertainty concepts
   - balance of views
   - interest in- and availability for the project

   According to EMSE280 (2011), during expert selection, the facilitator must beware the following:

   - pre-existing mindsets or assumptions that the expert may use
   - structural biases from the level of detail or choice of background scales for quantification
   - ulterior motives the experts may have for the study outcome
   - cognitive biases such as over-confidence, subconscious anchoring, and over- or underestimating due to how easily the experts recall events

7. Definition of elicitation format document

   The elicitation format document should provide an exact description of the questions and any explanation of them where necessary. It should also include the preferred formats for the uncertainty assessments to be provided by the experts and any additional remarks on what should be included in the assessments.

8. Dry run exercise

   This exercise aims to highlight anything ambiguous in the outlines of the elicitation format document, and to ensure that all relevant information and questions have been captured. One or two experienced people in the field of interest should provide comments on both the case structure and the elicitation format document before they are finalised and handed out to the panel of experts.

9. Expert training session

   To provide subjective assessments on the query and seed variables the experts provide subjective cumulative distributions in terms of 5%, 50% and 95% quantiles to indicate their degree of belief. Since most experts are not familiar with providing quantiles as a method of elicitation, they need to be trained in issues relating to subjective assessments.

Elicitation

10. Expert elicitation session

    Individual interviews of the experts are carried out with a normative analyst and a substantive expert present. In some instances the experts all meet together to discuss their assessments in a non quantitative manner, having had a chance to look at other experts assessments before-hand.

Post-elicitation

11. Combination of expert assessments

    Pooling the expert assessments is executed with the use of a software package called EXCALIBUR that supports three combination methods; equal, global and item weighting; which are discussed later in this section under "The Classical Model". The results indicate the facilitators probability distributions of the query variables.

12. Discrepancy and robustness analysis

    Robustness is analysed on the expert and seed variables in this step. The variables are removed one at a time and the target variable is re-calculated to determine the relative information loss due to the removed variable. A large loss means that one cannot replicate the results in another study with different variables. The uncertainty assessments where the experts differ most are identified using discrepancy analysis, then reviewed to determine whether any of the causes of the discrepancy are avoidable.

13. Feed-back

    Feed-back is normally done anonymously while every expert has access to his/her assessment, weights, calibration and informativeness score, and any passages where his/her name is used.

14. Post-processing analyses

    Further processing of the combined results is done to derive uncertainty distributions over the required input parameters using dedicated software packages.

15. Documentation

    The last step involves noting down all the relevant information and data in a formal report to be presented to the decision maker/client and to the experts.

The uncertainty spreads produced by the Cooke method are generally narrower than those produced by other approaches but wider than those of single experts. Aspinall (2010) highlights that the biggest challenge with this method is how to convey the resultant scientific uncertainty to the policy makers concerned as they would be expecting a sure result without any ambiguity. It is therefore of vital importance to figure out a useful method of communication to match the usefulness of having captured the uncertainties. It is also warned that the facilitator of this process (preferably a trained uncertainty analyst according to Cooke and Goosens (2010)) needs to be impartial and able to tactfully hinder individual experts from taking a pedestal during the debate. The facilitator also needs to be careful not to ask ambiguous or leading questions while maintaining the engagement of all the participants. If the expert panel consists of academics and practical field experts, ambiguity could lead to a separation between them due to different backgrounds. It would be important to highlight these differences from the beginning so that any errors or misunderstandings that arise may trigger further investigation rather than discourse.

It may also be a challenge to gather a good variety of useful experts as some scientists are reluctant to take part in societal matters, and some facilitators prefer to ignore uncertainty as it may result in loss of public confidence. Aspinall (2010) reassures that the Cooke method attempts to bridge these two camps in a more sensible way than the usual retreats to merely “agreeing to disagree” or to applying the the “precautionary principle” (a discretionary technique that leaves the burden of justifying a certain action to the implementor of the action when there can be no scientific basis for the decision).

According to Aspinall (2010), the longest running application of the Cooke method within science is volcano management in Montserrat strands.

The Classical Model

Cooke and Goosens (2010) define the Classical model as a “perfomance-based linear pooling or weighted averaging model”. The model begets its name from an analogy between classical statistical hypothesis testing and calibration measurement, and it contrasts with some Bayesian models.

Calculation of the weights is based on how the experts performed on the seed questions. The classical model contains three different performance-based weighting schemes for combining expert assessments: equal-, global- and item weighting. Cooke and Goosens (2010) explain that equal weighting assigns equal weights to each expert on the panel such that, if there are N experts, then each density has a weight of $\frac{1}{N}$, and the facilitator’s cumulative distribution function for variable i is given by

$$F_{ewdm,i} = \frac{1}{N}\sum_{j=1}^{N} f_{j,i}, \tag{2.8}$$

where $f_{j,i}$ is the cumulative probability associated with expert j's assessment of variable i, that is, $f_{j,i} = P_j(X_i \leq a)$. Global weights are determined, per expert, by the expert's calibration score from the seed variables, and the overall information score based on the uncertainty interval given by the expert. Item weights are also determined by the expert's calibration score, but these are calculated per expert and per variable to maintain regard for how well the experts are informed on each variable.

To give a measure of performance, the performance-based weights make use of calibration and information. Cooke and Goosens (2010) explain that "calibration" is a probability measure of whether the experimental results of the seed questions and the experts' assessments are in alignment, and any divergence between the two is alluded to chance. The probability measure is found by means of a hypothesis test. A low probability is interpreted as the experts' assessments likely being incorrect. "Information" refers to how concentrated an expert's distribution is, in relation to some background measure selected by the facilitator/client. According to Cooke and Goosens (2008), the uniform and log-uniform background measures are normally used. According to Cooke and Goosens (2010), the overall combined score given by calibration and information has the following properties:

- Information merely modulates between more- and less well calibrated experts so calibration dominates over information;
- To avoid bias and maintain neutrality, an expert's maximal expected score is achieved by- and only by stating his or her true beliefs;
- Due to the previous property, calibration is scored with a cut-off point and if the score is below the cut-off then the expert is unweighted;
- Optimising the calibration and information performance of the combination determines the cut-off point.

Ryan et al. (2012) emphasise that an expert's final weights are given by the normalised product of the calibration score and the information score. Once the weights are assigned, the process goes on to address the actual parameters of interest by combining the experts' distributions for these parameters, taking into account their relative weights.

As with Bayesian models, Cooke and Goosens (2010) point out that a fundamental assumption of the Classical model is that experts' future performance can be judged on the basis of how they have performed in the past. The classical model also follows the Procedures Guide outlined by the Cooke method.

Calibration

When confronted with continuous uncertainties, Cooke and Goosens (2010) explain that the expert is required to specify his subjective distribution over the possible values of the quantity in question.

Normally 5%, 50% and 95% quantiles are requested and then distributions are fitted to these quantiles based on the calibration and information scores from the quantiles. Ryan et al. (2012) goes on to say that the calibration of an expert relates to how well the experts specify their quantiles. A well calibrated expert will have:

- 5% of the seed variable observed values below his/her 5^{th} percentile value (i.e below q_1);
- 45% of the seed variable observed values between his/her 5^{th} and 50^{th} percentile values (i.e between q_1 and q_2);
- 45% of the seed variables observed values between his her 50^{th} and 95^{th} percentile values (i.e between q_2 and q_3); and
- 5% of the seed variables observed values above his/her 95^{th} percentile values (i.e beyond q_3).

The specifications above are referred to respectively as intervals 1 to 4. A useful visual representation of these intervals is shown in figure 2.5:

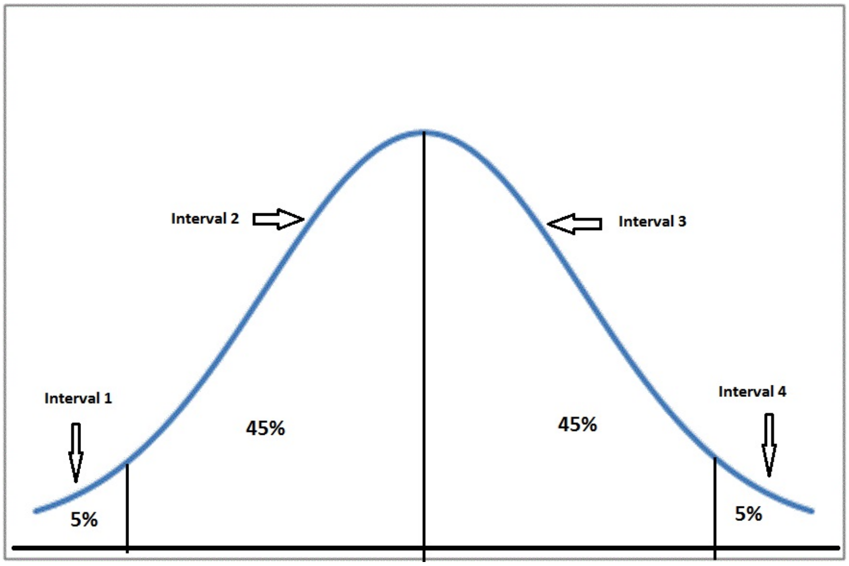


Figure 2.5: Calibration intervals

In discrete cases, Cooke and Goosens (2008) explain that the expert would be asked to assign each event to a pre-defined "probability bin" of $10\%, 20\%, ..., 90\%$. Cooke and Goosens (2008) favour the quantile format over the discrete format.

Ryan et al. (2012) explains that the calibration score is calculated using “relative information”, $I(s,p)$, given by

$$I(s,p) = \sum_{i=1}^{4} ln(s_i/p_i), \tag{2.9}$$

where s_i is $\frac{\text{the number of observed seed values in interval } i}{\text{total number of seed values}}$, and $p_i = 0.05, 0.45, 0.45$ and 0.05 from the intervals respectively. It is worth noting that if $s_i = p_i$, $\forall i$, then $I(s,p) = \sum_{i=1}^{4} ln(1) = 0$. Ryan et al. (2012) adds that when the number of seed variables, N, is large, $2NI(s,p)$ will have a $\chi^2_{(3)}$ distribution. The calibration score for each expert is then calculated as

$$cal = \begin{cases} P\{\chi^2_{(3)} > 2NI(s,p)\} & \text{if } P\{\chi^2_{(3)} > 2NI(s,p) > \alpha\} \\ 0 & \text{otherwise} \end{cases} . \tag{2.10}$$

This specifies a minimum calibration score of α that, if not met, means the expert receives zero weight for the analysis. If $s_i = p_i$, $\forall i$, then $I(s,p) = 0$ and $cal = P(\chi^2_{(3)} > 0) = 1$ which is the highest calibration score an expert can get.

According to Cooke and Goosens (2008), calibration scores can be compared across studies but one must first ensure that the power of the different hypothesis tests being compared is equalised. For instance, if the calibration scores of two data sets with N and N' realisations are being compared, then the minimum of N and N' is used.

Information

Ryan et al. (2012) explains that the entire expert distribution must be defined to be able to calculate the information score for each expert. In order to get the information score for expert e, where each seed variable’s known value will be r; the elicited percentile values are defined as $q_{5,e}$, $q_{50,e}$, and $q_{95,e}$. According to Eggstaff et al. (2014), for each item, consider the smallest interval $[L,U]$ to “bound the range of possible elicitation variables” . These bounds are defined as:

$$\begin{aligned} q_L &= min\{r, q_{5,1}, \ldots, q_{5,e}\}, \\ q_U &= max\{r, q_{95,1}, \ldots, q_{95,e}\}, \end{aligned} \tag{2.11}$$

where e is the number of experts.

Cooke and Goosens (2008) point out that uniform and log-uniform background measures make use of the “intrinsic range” where the measures are concentrated, which enables the classical method to implement the “k% overshoot rule”. According to Ryan et al. (2012), a k% (where k is normally set at 10 and chosen by the analyst) overshoot below q_L and above q_U for estimating the 0^{th} and 100^{th}

percentile points yields

$$\begin{aligned} q_0 &= q_5 - (k/100)(q_U - q_L) \text{ and} \\ q_{100} &= q_{95} + (k/100)(q_U - q_L). \end{aligned} \quad (2.12)$$

We then we have a full five point scale of q_0, q_5, q_{50}, q_{95} andq_{100}. Linear interpolation between the 5 points of q generates a complete expert distribution of the unknown quantity of interest. For quantities of interest that cannot be observed like the seed variables, the bounds are calculated as in equations 2.11 and 2.12, excluding r. The information score that indicates how the expert's distribution deviates from some meaningful background measure to reflect the expert's certainty in his/her answers is, again, calculated using relative information:

$$inf = I(h,p) = \sum_{i=1}^{4} p_i(p_i/h_i), \quad (2.13)$$

where

$$\begin{aligned} h_1 &= F_U(q_5) - F_U(q_0) = \frac{q_5 - q_0}{q_{100} - q_0}, \\ h_2 &= F_U(q_{50}) - F_U(q_5) = \frac{q_{50} - q_5}{q_{100} - q_0}, \\ h_3 &= F_U(q_{95}) - F_U(q_{50}) = \frac{q_{95} - q_{50}}{q_{100} - q_0}, \\ h_4 &= F_U(q_{100}) - F_U(q_{95}) = \frac{q_{100} - q_{95}}{q_{100} - q_0}, \end{aligned} \quad (2.14)$$

and F_U is the cumulative distribution function of the background measure. When the realisation is expected to fall between 0.001 and 1 000, Cooke (1991) recommends using the uniform distribution as the background measure, otherwise the log-uniform distribution.

Assuming the variables are independent, Cooke and Goosens (2008) explain that equation 2.13 is proportional to the joint distribution of the expert given the background. Information scores cannot be compared across studies because the scores depend on the "intrinsic range" and other experts' assessments.

Cooke and Goosens (2008) describe the relative information function as a slow function, such that if the expert assessments were to significantly change, the information score would be mildly affected. Conversely, the likelihood function of the calibration score is described as a fast function, which means the product of calibration and information is mostly driven by the calibration score. A high statistical likelihood/p-value and high information indicate "good expertise".

Decision Maker/Combination

Cooke and Goosens (2008) define a "decision maker" (*DM*) as the combination of expert assessments. The classical model derives weights for this combination process using linear pooling, and aims to select weights that result in "good expertise". A constraint on how the weights are derived is the "strictly scoring rule" that ensures that the weights are a form of reward to the expert for stating judgments that conform to his/her true beliefs in order to attain maximal weights. To obtain the weighting score for an expert *e*, the following formula is suggested:

$$w_\alpha(e) = 1_\alpha(calibration score) \times calibration score(e) \times information score(e), \tag{2.15}$$

where $1_\alpha(x) = 0$ if $x < \alpha$ and $1_\alpha(x) = 1$ otherwise; and $calibration score(e)$ and $information score(e)$ are calulated as in equations 2.10 and 2.13, respectively. This means a set of assessments are scored based on a set of realisations. α is chosen such that the resultant score of the *DM* is maximised. Using weights proportional to equation 2.15 and linear pooling for item *i*, the combined score is given by

$$DM_\alpha(i) = \frac{\sum\limits_{e=1,\ldots,E} w_\alpha(e) f_{e.i}}{\sum\limits_{e=1,\ldots,E} w_\alpha(e)}. \tag{2.16}$$

According to Cooke and Goosens (2008), if we let α^* maximise

$calibration score(DM_\alpha) \times information score(DM_\alpha)$ we obtain the global weight that is calculated on all seed items, DM_{α^*}. They add that one could use weights that employ individual information scores for each item, rather than the average information score, using the formula:

$$w_\alpha(e,i) = 1_\alpha(calibration score) \times calibration score(e) \times I(f_{e,i} \mid g_i), \tag{2.17}$$

where $f_{e,i} \mid g_i$ is expert *e*'s density for item *i*, given its background density g_i. Consequently, the resultant combined score is given by:

$$IDM_\alpha(i) = \frac{\sum\limits_{e=1,\ldots,E} w_\alpha(e,i) f_{e.i}}{\sum\limits_{e=1,\ldots,E} w_\alpha(e,i)}. \tag{2.18}$$

IDM_{α^*} is referred to by Cooke and Goosens (2008) as the item weight *DM* where α^* maximises $calibration score(IDM_\alpha) \times information score(IDM_\alpha)$. When experts are better trained in probabilistic assessment, item weights tend to improve over global weights. It is also explained that ideally, although not yet mathematically proven, the decision maker should perform better than the equal weight *DM* and the best expert on the panel.

The practical application of this popular classical model is illustrated through use of a simple example in chapter 3.

2.2.5 The Bayesian Approach

In this approach, Garisch (1985) explains that r experts with prior opinions about the other $r-1$ experts in the group are considered, and these opinions are used to formulate a posterior distribution for each expert. After this first round of assessments, the experts are considered to then posess individual subjective probability distributions. If there is a possibility of new information arising from a second round of assessments then the analyst may consider executing one. This revision is usually unnecessary as the experts would not be willing to compromise their opinions any further. It is not very probable that consensus will be reached after just the first round, but if consesnsus was a goal of the analysis then more revisions would be necessary. Since the experts are not likely to have a single estimate at any revision, the definition of "consensus" in this scenario becomes the revision with the minimum distance between the smallest and the largest estimates of a parameter.

Garisch (1985) mentions that, in order to maintain consistency, if an expert deems it necessary to change his assessment based on his opinion of all the other experts in the group, then on subsequent revisions he/she could be argued to have to also change his assessment based on his own opinion too. Unfortunately this argument does not lead to convergence for many revisions. It is suggested that it would be more rational for an expert to update his/her assessment of other people's estimates based simply on the expert's original opinion of them and the expert's knowledge and understanding of how his/her estimate will be used in their revisions.

If the more rational approach is followed, it is interesting to try and see whether there would be convergence to some mutual limit. Garisch (1985) explains that since each expert's estimate converges monotonically to an unbiased variation of the other experts' original opinions, there will not be any convergence to a mutual limit.

Decision-making technique

Suppose we have two experts who need to estimate some unknown parameter θ and must come as close to consensus as they can. Let expert i's estimate be denoted by θ_i, and let θ_i be the condition of a normal density function for his opinion:

$$\theta|E_i \sim N(\theta_i, \sigma_i) \quad i = 1,2$$

θ in this case is a random variable, and expert i's opinion about its variability is represented by σ_i.

Now, let expert i's opinion of expert j's expertise be represented as a normal density function on an estimate from expert j, θ_j :

$$\theta_j|E_i \sim N(p_i(\theta), w_i(\theta)) \quad i = 1,2$$

We will consider only the following particular functions of $p(\theta)$ and $w(\theta)$:

$$\begin{aligned} p_i(\theta) &= a_i + b_i\theta \\ w_i(\theta) &= v_i \quad i = 1,2 \end{aligned}$$

If we let $a_i = 0$ and $b_i = 1$ so that $p_i(\theta) = \theta$, we can then infer that expert i considered the estimate given by expert j to be an unbiased estimate of θ. Following receiving expert j's estimate, expert i must be ready to update his assessment about θ by forming a posterior opinion, $h_i(\theta|\theta_j)$, and using its mean as the new estimate.

Assuming independence of θ_1 and θ_2, we have that:

$$\theta|\theta_j, E_i \sim N\left(k_i\theta_i + (1-k_i)\frac{\theta_j - a_i}{b_i}, k_i\sigma_i\right), \tag{2.19}$$

where $k_i = \frac{v_i}{b_i^2\sigma_i + v_i}$. Expert i's new estimate will be given by:

$$\theta_i^{(1)} = k_i\theta_i + (1-k_i)\frac{\theta_j - a_i}{b_i} \quad i = 1,2$$

At this stage, each expert will have updated his estimate but consensus is unlikely as their updated estimates may consequently be even further apart. Consensus would have to be reached via compromise in later revisions. It is for this reason that the first revision is considered an "improved estimator", while any revisions thereafter are denoted "compromise estimators".

To illustrate this process of decision-making, consider a case of two experts, E_1 and E_2. Suppose their estimates of the unknown parameter θ are normally distributed as follows:

$$\begin{aligned} \theta|E_1 &\sim N(\theta_1 = 10, \sigma_1 = 2) \text{ and} \\ \theta|E_2 &\sim N(\theta_2 = 15, \sigma_2 = 5). \end{aligned}$$

Suppose further that expert i's opinion of expert j's expertise are given by:

$$\begin{aligned} \theta_2|E_1 &\sim N(p_1(\theta) = a_1 + b_1\theta = \theta_1 + 7\,, w_1(\theta) = v_1 = 5) \text{ and} \\ \theta_1|E_2 &\sim N(p_2(\theta) = a_2 + b_2\theta = \theta_2 + 3\,, w_2(\theta) = v_2 = 4). \end{aligned}$$

It follows that for $\theta_2|E_1$, $a_1 = 7$, $b_1 = 1$ and $v_1 = 5$; and for $\theta_1|E_2$, $a_2 = 3$, $b_2 = 1$ and $v_2 = 4$. Assuming independence of θ_1 and θ_2, we performed 4 revisions on the original values of θ. As per equation 2.19, let k_E_i represent the estimate of k_i for expert i, θ_E_i represent the updated estimate of θ_i for expert i, and σ_E_1 represent the updated estimate of σ_i for expert i. We obtained the following results:

Repetition	k_E_1	θ_E_1	σ_E_1	k_E_2	θ_E_2	σ_E_2	$[(\theta\_E_1)-(\theta\_E_2)]^2$
1	-	10	2	-	15	5	-
2	0.714286	9.428571	1.428571	0.444444	10.555556	2.222222	1.270093
3	0.777778	8.123457	1.111111	0.642857	9.081633	1.428571	0.918101
4	0.818182	7.024943	0.909091	0.736842	8.040007	1.052632	1.030355
5	0.846154	6.104184	0.769231	0.791667	7.203536	0.833333	1.208574
6	0.866667	5.317431	0.666667	0.827586	6.496751	0.689655	1.390796
7	0.882353	4.632645	0.588235	0.852941	5.882145	0.588235	1.561250
8	0.894737	4.027329	0.526316	0.871795	5.337337	0.512821	1.716121

Table 2.1: Decision-making technique results

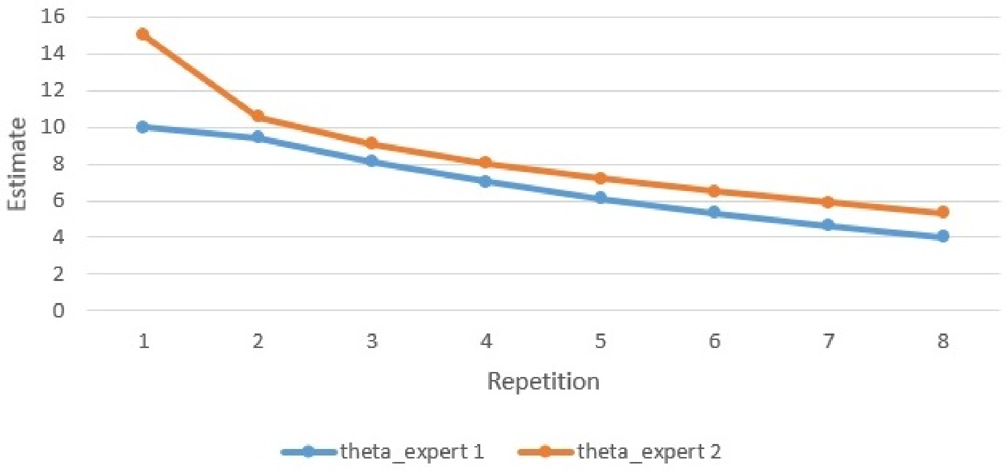


Figure 2.6: Decision-making technique

As shown by table 2.1 and figure 2.6, the initial distance between the experts' estimates was 5. The smallest squared distance of 0.918101 was achieved at the first revision of their estimates, thus, these are referred to as the "improved estimators". The revisions thereafter are denoted "compromise estimators" as their aim was to determine if consensus could actually be achieved and it was not, thus confirming the notion that consensus is highly unlikely.

Now suppose that the experts reveal their revised estimates to each other. The opinion they have of each other's estimates is based on the original opinion and updated adhering to Bayes rules. If the experts still don't achieve consensus, then they continue compromising until they get as close as possible to consensus. After n revisions, expert i will have an estimate:

$$\theta_i^{(n)} = \frac{\left\{\theta_i k_i v_i + \theta_j\left(\frac{(1-k_i)v_i}{b_i} + (n-1)k_i\sigma_i b_i\right) - \frac{v_i(1-k_i)a_i}{b_i} - (n-1)k_i\sigma_i a_i b_i\right\}}{\left\{v_i + (n-1)k_i\sigma_i b_i^2\right\}},$$

with variance $\sigma_i^{(n)} = \frac{k_i\sigma_i v_i}{\left\{v_i+(n-1)k_i\sigma_i b_i^2\right\}}$.

The decision making technique described above for two experts can be extended to a group of r experts.

It is suggested that the revision where the estimates are closest to each other should be determined instead of trying to reach consensus. This will be the revision where the "square of the difference between the maximum and minimum estimates is a minimum". One could also obtain the posterior distributions at the revision where the estimates are closest to each other, calculate quantiles, and then proceed as previously described in the classical model in section 2.2.4.

2.2.6 Using the Dirichlet distribution

Cooke and Goosens (2008) explain that many believe that expert judgment should be approached from the Bayesian paradigm since experts' uncertainties concern their subjective probabilities. This paradigm is based on achieving the maximal expected utility for each individual by requesting a joint distribution that takes into account the variable of interest, the seed variables and the expert's distributions over these quantities. Cooke and Goosens (2008) argue that "rational consensus" issues are resistant to the Bayesian approach.

According to Zapata-Vazqeuz et al. (2012), expert knowledge is increasingly being confidently represented as a probability distribution but, eliciting multivariate distributions is a more complex task as it does not suffice to assume mutual independence of the quantities and elicit their marginal distributions separately. An even further challenge is to elicit expert judgments about possible correlations between uncertain quantities. The Dirichlet distribution is applicable in multivariate elicitation tasks where the facilitator is trying to elicit information about a set of proportions that must add up to 1. Furthermore, the Dirichlet is a conjugate to multinomial likelihood, which makes it a good choice to elicit prior distributions for use in Bayesian analysis. The use of the Dirichlet distribution in maintenance optimisation by Van Noortwijk et al. (1992) will be discussed in section 2.2.6.3.

2.2.6.1 Properties of the Dirichlet distribution

According to Zapata-Vazqeuz et al. (2012), an expert's belief π has a Dirichlet distribution with parameter vector $\mathbf{d} = (d_1, ..., d_k)$ for k uncertain quantities, i.e $\pi \sim Di(\mathbf{d})$ if the belief has probability density function

$$f(\pi \mid \mathbf{d}) = c(\mathbf{d}) \prod_{i=1}^{k} \pi_i^{d_i - 1}, \tag{2.20}$$

where $\pi_i \geq 0$ for $i = 1, 2, ..., k$, $\sum_{i=1}^{k} \pi_i = 1$ and the normalising constant is $c(d) = \frac{\Gamma\left(\sum_{i=1}^{k} d_i\right)}{\prod_{i=1}^{k} (\Gamma d_i)}$.

Given that $n = \sum_{i=1}^{k} d_i$, the mean, variance and covariance of the π_i's are, respectively, given by

$$\begin{aligned} E(\pi_i \mid d_i) &= \frac{d_i}{n}, \\ Var(\pi_i \mid d_i) &= \frac{d_i(n-d_i)}{n^2(n+1)}, \\ Cov(\pi_i, \pi_j \mid d_i) &= \frac{d_i d_j}{n^2(n+1)}. \end{aligned} \tag{2.21}$$

Zapata-Vazqeuz et al. (2012) point out that the Dirichlet distribution properties useful for elicitation are the marginal and conditional properties.

The marginal property is that $\pi^{m+} \sim Di(d_1, d_2, ..., d_m, d^*_{m+1})$ where $d^*_{m+1} = \sum_{i=m+1}^{k} d_i = n - \sum_{i=1}^{m} d_i$. This is due to the first m elements of π being denoted by $\pi^m = (\pi_1, \pi_2, ..., \pi_m)$, and

$\pi^*_{m+1} = \sum_{i=m+1}^{k} \pi_i = 1 - \sum_{i=1}^{m} \pi_i$, such that $\pi^{m+} = (\pi^m, \pi^*_{m+1}) = (\pi_1, \pi_2, ..., \pi_m, \pi^*_{m+1})$ and π^{m+} lies on the $m-$ dimensional simplex. In addition, there exists a special case where the marginal distribution of π_i is a beta distribution with parameters d_i and $n - d_i$.

For the conditional property let $\pi'_i = \frac{\pi_i}{(1-\pi_1-\pi_2-...-\pi_m)}$ for $i = m+1, ..., k$, so $\pi' = (\pi'_{m+1}, ..., \pi'_k)$ lies on the $(k-m-1)$ dimensional simplex. The conditional property is thus that

$\pi' \mid \pi^m (\text{or}\, \pi^{m+}) \sim Di(d_{m+1}, ..., d_k)$. Zapata-Vazqeuz et al. (2012) advise that repeated use of this property enables one to decompose the Dirichlet distribution into $k-1$ beta conditional distributions.

As a simple demonstration, the opinions of one expert on a set of k uncertain quantities:

$\pi = (\pi_1, \pi_2, ..., \pi_k)$ can be elicited. These uncertainties lie on a $(k-1)$ dimensional simplex where $\pi_i \geq 0$ for $i = 1, 2, ..., k$ and $\sum_{i=1}^{k} \pi_i = 1$, enabling the use of the Dirichlet family. The result of the elicitation is expected to be a representation of the expert's knowledge about π as a probability distribution (expected to also be of the Dirichlet family) on the simplex. Zapata-Vazqeuz et al. (2012) further explain that the Dirichlet family is sometimes parameterised with $p_i = \frac{d_i}{n}$ $(i = 1, 2, ..., k)$ and n, instead of the k parameters d_i. To ensure that there are k parameters, there is a constraint that $\sum_{i=1}^{k} p_i = 1$. The p_i's control the means of the π_i's while n controls the overall quantity of uncertainty, thus the expert is requested to provide an opinion of these parameters. Literature proposes various ways of eliciting p_i's using the Dirichlet distribution such as requesting the median or mode based on the assumption of equal probability, requesting a joint mode of the predictive distribution, or using simple probabilities assuming the i's sum to 1. A value for n can be elicited by obtaining a measure of uncertainty for one π_i, or by considering n to be a measure of the amount (or strength) of information an expert has.

Zapata-Vazqeuz et al. (2012) mention that the fact that the Dirichlet distribution has only one parameter, n, to control the overall uncertainty is a drawback. If an expert's uncertainty about different π_i's did not quite agree with the constraints surrounding the choice of n then a Dirichlet distribution would not work well. Employing a mixture of Dirichlet distributions is suggested, but this would increase the number of parameters, and consequently increase the number of judgments to be elicited.

2.2.6.2 Using the Dirichlet distribution in multivariate elicitation for a set of proportions that must sum up to 1

When the afore methods of elicitation using the Dirichlet distribution were suggested, Bayesian analysis was limited in that the prior distribution had to be a member of the relevant conjugate family. Zapata-Vazqeuz et al. (2012) explain that it is now possible to elicit expert beliefs about proportions without necessarily aiming to elicit a Dirichlet distribution, i.e. the outcome would either be that the Dirichlet distribution is an acceptable representation of an expert's opinion, or that no Dirichlet distribution would adequately represent the expert's opinion. Taking these recent developments into account, the approach by Zapata-Vazqeuz et al. (2012) employs a technique known as "over-fitting". This means that when fitting the Dirichlet distribution, they elicit more judgments than necessary from the expert. Having more elicited quantities allows the facilitator to better identify a suitable choice of **d**. Over-fitting also allows the facilitator to assess the degree to which the Dirichlet distribution will represent an expert's judgments with a well chosen **d**, and thus decide whether the Dirichlet is adequate or not. A package that employs over-fitting known as the Sheffield Elicitation Framework (SHELF) was used to elicit a distribution for each expert judgment π_i by finding the best n and assessing how adequately the Dirichlet fits. Briefly, this procedure is as follows:

1. Preparation and training

   This phase involves getting familiar with SHELF and training the experts on the elicitation method that will be used and how to give personal probability judgments and distributions.

2. Eliciting beta distributions for each π_i using the SHELF package

   This step is executed for each experts belief, π_i, in turn. While incorporating over-fitting, SHELF identifies the parameters of the respective beta distributions for the experts' beliefs. The facilitator then summarises the results and gives feedback to the expert, who decides whether their beliefs have been adequately represented. If the software failed to adequately represent their π_i's using beta distributions then the process is terminated and it is reported that no Dirichlet distribution can represent the expert's judgments.

3. Checking and adjusting the means

   If the process does not terminate at step 2, the means are then adjusted to conform to the constraint that $\sum_{i=1}^{k} \pi_i = 1$. For instance, if it is found that $\pi_i \sim Beta(d_i, e_i)$, then $E(\pi_i) = \frac{d_i}{d_i+e_i}$, so we must have that

   $$\sum_{i=1}^{k} \frac{d_i}{d_i + e_i} = 1. \tag{2.22}$$

   It is unlikely to actually obtain d_i and e_i values that satisfy equation 2.22, so these quantities often have to be adjusted as long as they fall in the range $[0.9, 1.1]$, otherwise the whole elicitation has to be reviewed to identify where the expert may have misunderstood something. The sum on the left hand side of equation 2.22 is denoted by r and as long as $r \neq 1$, then d_i and e_i are adjusted as follows:

   $$d_i^* = \frac{d_i}{r}, \quad e_i^* = d_i + e_i - d_i^*. \tag{2.23}$$

4. Finding the best choice of n

   Ideally, if the values of $n_i = d_i^* + e_i^*$ are all equal then the separate beta distributions correspond to a Dirichlet distribution. In practice, these n_i values are disparate so a compromise n value is found and the Dirichlet distribution can be defined by

   $$Di(\mathbf{d}(n)), \tag{2.24}$$

   where the i^{th} element of $\mathbf{d}(n)$ is given by

   $$d_i(n) = n \frac{d_i^*}{d_i^* + e_i^*}.$$

   Choices of the optimal n_{opt} include the following:

   - A "compromised" value, i.e a strict middle value $n_{mid} = \frac{(n_{min}+n_{max})}{2}$, or the mean $\bar{n} = \sum \frac{n_i}{k}$, or the median n_{med}.

- A "simplified optimisation", i.e a value for n that optimises some criterion $F(n)$ that is based on the extent to which equation 2.24 matches the standard deviations of the individual beta distributions elicited in step 2. The optimised value is given by

$$n_{opt} = \left(\frac{\sum_{i=1}^{k} v_i^* (n_i + 1)}{\sum_{i=1}^{k} v_i^* \sqrt{n_i + 1}} \right)^2 - 1, \tag{2.25}$$

  where $v_i^* = \frac{d_i^*(n_i - d_i^*)}{n_i^2(n_i+1)}$ is the variance of the i^{th} adjusted beta distribution $Beta(d_i^*, n_i - d_i^*)$, using equation 2.21.

- A "conservative" value, i.e $n = n_{min}$ as one does not presume knowing any more about any quantity, π_i, than was elicited.

5. Providing feedback to the experts

   Once the best possible n is computed, the expert is made aware of the implications of the fitted Dirichlet distribution so he/she can decide if it is acceptable before the process is terminated by concluding with the elicited distribution. In the event that the n_i values were not originally similar, then it is concluded that the Dirichlet will not adequately represent the expert's knowledge as the fit is now too poor.

2.2.6.3 Using the Dirichlet distribution in maintenance optimisation

Due to the costliness of downtime from failure on a production unit, Van Noortwijk et al. (1992) explain that it is important to figure out an optimal maintenance interval to avoid having to perform maintenance activities too frequently. According to Van Noortwijk et al. (1992), Konink-lijke/Shell Laboratorium in Amsterdam has been using a decision support system called PROMPT to get component lifetime distributions towards maintenance optimisation. The challenge is that failures aren't often observed for well maintained systems, so there is a lack of reliable data to determine the distributions, thus the need for the use of expert judgment analysis.

Van Noortwijk et al. (1992) initiated the use of expert judgment analysis within the maintenance environment, and they demonstrated their procedure using a single component. Their procedure involved the histogram technique, the Supra Bayes approach and use of the Dirichlet distribution. Van Noortwijk et al. (1992) specify that for elicitation, it is important to consider the nature of information required, the number of available experts, the background and training of the experts, the number of variables of interest and the speedy execution of the elicitation process. Since the facilitators needed to model how a component was aging, they had to use a time-based preventive maintenance policy.

The histogram technique

A challenge with elicitation is that experts have to fit an entire parametric life distribution which, according to Van Noortwijk et al. (1992), is difficult even for experts trained in statistics. For this reason, along with its simplicity, easy comprehension and good predictive probability, the histogram technique was chosen to be the basis of Van Noortwijk et al. (1992)'s elicitation procedure. Using such a discrete technique would mean the probability of failure can be displayed in fixed time intervals versus a probability density.

Suppose the domain of lifetimes for a product has a range $(0, \infty)$. Let the domain be divided into m disjoint time intervals so that we have $(t_{i-1}, t_i]$ for $i = 1, 2, ..., m-1$, till the m^{th} interval (t_{m-1}, ∞). Van Noortwijk et al. (1992) explain that since maintenance engineers replace most of the components before failure, they only have experience with the first part of the life-cycle, which necessitates the m^{th} interval being an open interval as it would be difficult to accurately report the tail of the lifetime cycle. Van Noortwijk et al. (1992) deem it important to also note that in implementing the histogram technique, it is tricky to decide on a selection of m, t_i and a method to obtain interval failure probabilities from the experts. This challenge results in "a trade-off between limiting the accuracy obtained through elicitation and requiring an accuracy from the expert which cannot be achieved". Placing high importance on making elicitation as simple a process as possible for the experts, compromise solutions were employed, such as choosing $m = 5$ as researched and advised by psychological experiments. It is also advisable to have the intervals equidistant for better visual perception, and furthermore, have them somewhat based on past maintenance intervals so that they are familiar to the maintenance expert. Since components are expected to fail in the last interval, the afore-mentioned approach will likely have a large tail that will also likely contain the optimal maintenance interval. Van Noortwijk et al. (1992) advise that if this happens to be the case, one would have to resort to a continuous distribution. Van Noortwijk et al. (1992) also report that using laymen terms such as "1 in a 100" (as opposed to 0.01) and allowing probabilities to be represented as percentages with $n = 100$ makes the elicitation process easier for the experts.

When this article was written in 1992, it was worth noting that a computer based program that sped up the process and gave instant on-line visual feedback accurately was key in facilitating the process with ease. Van Noortwijk et al. (1992) encourage the presence of a trained analyst to guide the expert while using the computer program, which would still be essential today.

The Supra Bayes approach

This approach was developed for the purposes of combining judgments from several experts and reaching a consensus, as explained by Van Noortwijk et al. (1992). The supra Bayes approach assesses expert opinion by analysing and incorporating background information, i.e. the experts' expertise, prediction ability and prior information sets. When specifying a prior distribution for a quantity of interest and deciding on a likelihood function for the experts' judgment, this background information has to be taken into account. The Bayes theorem is used to combine the information in the form of a posterior distribution.

The multinomial case that arises with the use of the histogram technique requires a slight adjustment of the "natural conjugate" Bayes approach (where the prior distribution is a natural conjugate of the likelihood function to be used). Suppose in the case of maintenance optimisation, n components of the same type are installed at time t_0. The expert is asked to provide the number of components, n_{ie}, that he or she thinks will fail in time interval i. The probability of failure within interval i is therefore given by $p_{ie} = \frac{n_{ie}}{n}$, $i = 1,...,m$. Van Noortwijk et al. (1992) give each expert e's likelihood as

$$\mathscr{L}_e(\mathbf{p_e} \mid \mathbf{p_A}) = \prod_{i=1}^{m} (p_{iA})^{w_e \beta n p_{ie}} = \prod_{i=1}^{m} (p_{iA})^{w_e \beta n_{ie}}, \tag{2.26}$$

where $w_e \geq 0$ is the weight of each expert determined from their prior beliefs, β is a measure determined from the prior beliefs of all the experts as a group, $\mathbf{p_e}$ is a vector of the subjective probability distributions of failure specified by the experts, and $\mathbf{p_A}$ is a vector of the probability distributions of failure for the component. The exponent $w_e \beta n_{ie}$ is considered to be the number of virtual observations. Van Noortwijk et al. (1992) further define the likelihood for all available expert opinion as

$$\begin{aligned} \mathscr{L}(\mathbf{p_1},...,\mathbf{p_E} \mid \mathbf{p_A}) &= \prod_{e=1}^{E} \mathscr{L}_e(\mathbf{p_e} \mid \mathbf{p_A}) \\ &= \prod_{e=1}^{E} \left\{ \prod_{i=1}^{m} (p_{iA})^{w_e \beta n_{ie}} \right\} \\ &= \prod_{i=1}^{m} (p_{iA})^{\beta \sum\limits_{e=1}^{E} w_e n_{ie}}, \end{aligned} \tag{2.27}$$

where the exponent $\beta \sum\limits_{e=1}^{E} w_e n_{ie}$ now refers to the "number of virtual observations" for the experts as a group, meaning β controls the magnitude. According to Van Noortwijk et al. (1992), the natural conjugate for the multinomial likelihood, and also a well suited prior distribution for $\mathbf{P}_A$, is the Dirichlet distribution (as introduced by equation 2.20) and in this case it is given by

$$\pi(p_{1A},...,p_{m-1A}\mid\alpha)=\frac{\Gamma(\alpha_0)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m}p_{iA}^{\alpha_i-1}, \tag{2.28}$$

where $\alpha_i>0$, $\sum_{i=1}^{m}\alpha_i=\alpha_0$ and $\sum_{i=1}^{m}p_{iA}=1$ for $i=1,...,m$. Van Noortwijk et al. (1992) add that prior information about the probability function could also be incorporated using the mean and variance, as introduced in equation 2.21, using the following adaptations:

$$\begin{aligned}E(P_{iA}) &= \frac{\alpha_i}{\alpha_0},\\ Var(P_{iA}) &= \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}.\end{aligned} \tag{2.29}$$

So if we let the decision maker's best guess about the true value of $\mathbf{P}_A$ be denoted by $\mathbf{p}_A^*$, Van Noortwijk et al. (1992) suggest that it may be incorporated into the prior distribution by letting $\alpha_i=p_{iA}^*\alpha_0$, where α_0 controls the variance of P_{iA} and indicates the decision maker's strength of his/her belief in his prior estimate of $\mathbf{P}_A$. The posterior distribution of $\mathbf{P}_A$ is then calculated as being proportional to the product of the likelihood in equation 2.27 and the prior distribution in equation 2.28, now with $\alpha_i=\beta\sum_{e=1}^{E}w_en_{ie}$. Since $\sum_{i=1}^{m}\alpha_i=\alpha_0$, and based on equation 2.29, the posterior mean is given by

$$\begin{aligned}E(P_{iA}\mid\mathbf{p}_1,...,\mathbf{p}_E) &= \frac{\beta\sum_{e=1}^{E}w_en_{ie}}{\sum_{i=1}^{m}\beta\sum_{e=1}^{E}w_en_{ie}}\\ &= \frac{\sum_{e=1}^{E}w_ep_{ie}}{\sum_{e=1}^{E}w_e},\end{aligned} \tag{2.30}$$

where $\sum_{i=1}^{m}n_{ie}=1$, and since the experts share, more or less, the same information the weights are also restricted as $\sum_{e=1}^{E}w_e=1$. Equation 2.30 helps to derive the posterior estimate of $\mathbf{P}_A$, which is also the decision maker's consensus distribution, $\mathbf{P}_D$.

To combine opinions from several experts, Van Noortwijk et al. (1992) used the Supra Bayes approach along with determining weights for the experts. To get the optimal set of weights they needed a solution that yielded the best distribution for the decision maker. Let $\mathbf{r} \equiv (r_1, ..., r_m)$ be the recorded number of failures of the component used to calibrate the experts in time interval i, where $i = 1, ..., m$; $\mathbf{c} \equiv (c_1, ..., c_m)$ be the recorded number of maintenance activities for the component used to calibrate the experts in time interval i, where $i = 1, ..., m$ and $c_m \equiv 0$; $p_{iD} = \sum_{e=1}^{E} w_e p_{ie}$ be the consensus probability distribution; and the likelihood of existing failure be given by

$$\mathscr{L}(\mathbf{r}, \mathbf{c} \mid \mathbf{p}_D) = \prod_{i=1}^{m} (p_{iD})^{r_i} \left(\sum_{j=i+1}^{m} p_{jD} \right)^{c_i}. \tag{2.31}$$

Now letting $\sum_{e=1}^{E} w_e = 1$, according to Van Noortwijk et al. (1992), the weights are found using

$$w = \underset{w}{ARGMAX} \left\{ \mathscr{L} \left(\mathbf{r}, \mathbf{c} \mid \sum_{e=1}^{E} w_e \mathbf{p}_e \right) \right\}. \tag{2.32}$$

Van Noortwijk et al. (1992) add that an alternative that is easy to obtain is using “normalised likelihood weights” given by

$$w_e = \frac{\mathscr{L}(\mathbf{r}, \mathbf{c} \mid \mathbf{p}_e)}{\sum_{f=1}^{E} \mathscr{L}(\mathbf{r}, \mathbf{c} \mid \mathbf{p}_f)} \quad e = 1, ..., E \tag{2.33}$$

as good starting weight values for optimisation.

Van Noortwijk et al. (1992) explain that posterior to expert opinion, the decision maker’s distribution’s measure of β controls the variability as α_0 did for P_{iA}. β indicates the extent of the decision maker’s belief in the consensus estimate and also “controls the sensitivity of the posterior distribution to the sample information”. Van Noortwijk et al. (1992) further elaborate that there are two ways of eliciting β from the decision maker. One way is to test the decision maker’s sensitivity to new information. Using equation 2.30, p_{jD} can be calculated for a fixed value j. The decision maker would then have to indicate how his estimate would change if some appropriately chosen value on n^* new failures were observed. α_0 would change to $\beta n + n^*$ and the new estimate of P_{jA} would become $p^*_{jD} = E(P_{jA} \mid \mathbf{p}_1, ..., \mathbf{p}_E, \mathbf{n}^*) = \left(\frac{\beta n}{\beta n + n^*}\right) p_{jD} + \left(\frac{n^*}{\beta n + n^*}\right)$. Then finally the decision maker would have to solve for $\beta = \frac{n^*(1 - \tilde{p}_{jD})}{n(\tilde{p}_{jD} - p_{jD})}$, where $\tilde{p}_{jD} = p^*_{jD}$. This is executed for several different values of j to evaluate the consistency of the decision maker. Van Noortwijk et al. (1992) describe another way to elicit β where the decision maker gives a range R_j for any P_{jA}, which can be equated to 6 standard deviations $R_j^2 = 6^2(Var(P_{jA}))$. Using equation 2.29, one can then solve for $\beta = \frac{1}{n}\left[\frac{36p_{jD}(1-p_{jD})}{R_j^2} - 1\right]$.

Finally, Van Noortwijk et al. (1992) explain how the Bayes theorem is used to update the decision maker's probability distribution $\mathbf{P}_D$ with new data. If we represent the new data as $\mathbf{s} = (s_1, ..., s_m)$ failures and $\mathbf{v} = (v_1, ..., v_m)$ maintenance actions where $v_m \equiv 0$, then the likelihood is given by

$$\mathscr{L}(\mathbf{r}, \mathbf{c} \mid \mathbf{p}_A) = \prod_{i=1}^{m} (p_{iA})^{s_i} \left(\sum_{j=i+1}^{m} p_{jA} \right)^{v_i} . \tag{2.34}$$

The posterior distribution is then calculated as being proportional to the product of the likelihood in equation 2.34 and the prior distribution in equation 2.28 with $\alpha_i = \beta \sum_{e=1}^{E} w_e n_{ie}$, giving a "Generalised Dirichlet distribution":

$$\prod_{i=1}^{m} p_{iA}^{\alpha_i + s_i - 1} \left(\sum_{j=i+1}^{m} p_{jA} \right)^{v_i} , \tag{2.35}$$

with the posterior mean:

$$\begin{aligned} E(P_{iA} \mid \mathbf{p}_1, ..., \mathbf{p}_E, s, v) &= (s_i + \alpha_i) \\ &= \frac{\prod_{h=1}^{i-1} \left[\sum_{z=h}^{m-1} (\alpha_{z+1} + s_{z+1} + v_z \right]}{\prod_{h=1}^{i} \left[\sum_{z=h}^{m} (\alpha_z + s_z + v_z \right]} \quad i = 1, ..., m. \end{aligned} \tag{2.36}$$

Equation 2.36 is the updated probability distribution of the decision maker, in light of new data. Van Noortwijk et al. (1992) further advise that the covariance indicates the smoothness of the posterior distribution, while the variances are useful for determining bounds for the distribution.

2.3 Some practical uses of Expert Judgment to date

- In 2003, Aspinall (2010) put together a group of experts to discuss how water leaks into aging dams and eventually causes failure. The experts each gave their own estimation of the time a specific dam would take to fail once a leakage had started. Along with this estimate, to quantify its uncertainty, each expert gave a "credible interval" that gave their true answer a lee-way of 10%. Confident experts provided very narrow ranges while the more cautious experts provided longer time estimates with wider ranges of uncertainty. This is an example of the cognitive bias mentioned in section 2.2.4.

  Each expert then had to answer 11 seed questions in order to have their proficiency calibrated.

Their opinions were weighted based on how they performed on the seed questions and these weighted opinions were then combined to give a “rational consensus”. It turned out that the more cautious experts performed better on the seed questions resulting in their estimates and their intervals being weighted more heavily. Aspinall (2010) points out that this was one of the advantages of using the Cooke method: these cautious experts would have been poorly represented if the more prominent experts were allowed to steer the decision, but this method encourages these experts who would normally be wary of participating in policy advice by relieving them of the “burden of sole responsibility” and reassuring them with the use of a structured, neutral and collective procedure.

- Ryan et al. (2012) used expert judgment in the field of information security over a period of 6 months to attempt to determine:

  - how often a computer or system comes under attack;
  - how many of the attacks are successful;
  - whether it is worthwhile to invest a lot into protecting a system from attacks; and
  - the probability of a successful attack under different scenarios.

  Experts were asked 31 questions, 10 of which were seed variable questions developed by information security specialists knowing that the experts would not have direct access to the actual answers but should have adequate knowledge to make the required assessments. Experts were given IDs to preserve anonymity and their calibration and information scores were determined based on their responses to the seed variable questions. The expert judgment analysis employed the classical model through Microsoft Excel. As Cooke and Goosens (2010) suggested, the calibration scores had a more pronounced range than that of the information scores since calibration is more important.

  The 31 questions included sections where the same set of questions was posed in 3 different scenarios. The expert distributions for questions in these sections that referred to external behaviour unaffected by a change in scenario were identical distributions. It was also observed that distances between distributions from a scenario with security were greater than in others, indicating that there are benefits from some protection.

  The aim of this research was to show the development of prior distributions for the parameters of Non-Homogeneous Poisson Process (NHPP) models for cyber attacks and the expert judgment results were going to be used to develop distributions for both HPP and NHPP forms towards this. The instrument for eliciting the expert judgments was developed and executed by two information security specialists and two expert judgment analysis specialists. The resultant instrument was then validated by a small set of information security experts.

- In the increasingly important task of assessing the risk of wire failure in modern aircraft, Mazzuchi et al. (2008) explain how the paired-comparison technique for expert judgment was used to model the relationship for the probability of wire failure as a function of influencing factors in the zonal environment of the aircraft. Expert judgment analysis was appropriated by the lack of historical data in the field of wire failure, so a paired comparison workshop using 14 experts from the aviation community was held to obtain relative failure rates of different environments and ultimately express wire failure rate as a function of wiring environment.

  The pdf for how long it takes wires to fail for two independent failure modes, "fail to ground" and "fail to open", is taken to be exponential and given by

  $$f(t_i|\lambda_i) = \lambda_i e^{-\lambda_{it}}, \tag{2.37}$$

  where the two modes are represented by $i = g,o$ and $\lambda_i > 0$ is the failure rate that is usually calculated from past data for failure mode i. Often the time to failure may be accelerated by environmental factors and the proportional hazards model (PHM) was employed to incorporate variables that represented the 13 environmental factors present. The failure rate in equation 2.37 was expressed as a function of the covariates using the form $\lambda = e^{\beta_0+\beta_1 X_1+\ldots+\beta_{13}X_{13}}$. The resultant pdf was then represented as:

  $$f(t|\beta_0,\ldots,\beta_{13}) = e^{\beta_0+\Sigma_{j=1}^{13}\beta_j X_j} \times exp\{-[e^{\beta_0+\Sigma_{j=1}^{13}\beta_j X_j}]t\}, \tag{2.38}$$

  where t represents the "time to failure", X_j represents the quantitative effect of covariate j and, β_j (which must be estimated from past data) represents regression parameters relating the influence of covariate j on the failure rate.

  The ratio of the total number of observed failures to the total exposure time is a common estimate of failure rate. The construction of this estimate requires data from all possible failure environments which is overwhelmingly impractical, and results in a lack of data to support such an inference, thus the use of expert judgment/subjective data. This research made use of the NEL model (based on the paired-comparison method) by doing the following:

  - Selecting a number of failure environments (of which at least one had a reasonable amount of already existing failure data) to compare using the paired-comparison method. 15 sample environments were selected, without expert input, to be used in the elicitation. The environment selection was based on realism, change in environment comparisons, and a wide coverage of possible wiring environments. The experts were asked to fill out a survey of 105 questions that compared different environments and they had to rank the environments in each question. While being analysed for individual and group performance, 5 experts were removed for failing the test for consistency. The aim was to have the result of this process be a set of failure rate estimates obtained to within a proportionality constant;

- The failure rate estimates were obtained using a computer program called WCOMPAR and combined with their joint 90% bounds.The intention was to use these failure rate estimates to determine parameter estimates of $\beta_0, \dots, \beta_{13}$.
- A regression analysis was performed on the compared environments and the aim was to then obtain a failure rate for the candidate environment that had significant exposure time and failure data, but the process was not completed;
- The constant of proportionality for all failure rate estimates was then to be calculated using the results of the previous two steps.

Once all the parameters were estimated, it was then possible to specify the complete failure rate and corresponding pdf for any environment.

In conclusion, the procedure was a useful starting point for thorough risk analysis of any failure environment and it was highlighted that the procedure produced estimates and not truths, as it is still a new procedure to the field of aviation and observed data would be needed for proofing purposes.

- To facilitate a multivariate elicitation process, Zapata-Vazqeuz et al. (2012) used an extension of the SHELF package in a new method that employs over-fitting to produce either a more carefully considered Dirichlet distribution or an alternative resolution that the Dirichlet would not be a reasonable fit. Zapata-Vazqeuz et al. (2012)'s exercise was assessing the effectiveness of a new antibiotic on pediatric pneumococci patients. After the administration, the possible outcomes were: to survive in good condition (π_1), to have a sequel (π_2) or to die (π_3); and these proportions had to sum up to 1. As per the procedure outlined in section 2.2.6, Zapata-Vazqeuz et al. (2012) proceeded as follows:

    1. Preparation and training

       The facilitator decided to use SHELF's Quartile method for its applicable protocols on eliciting information about a single distribution, and installed *rpanel* so he/she could use SHELF's R functions.

       One expert was trained in the purposes of elicitation and how to execute probability judgments, then given a practice exercise on the Quartile method. The expert was also given background information on the nature of the antibiotic and the resulting categorisation of the patients.

2. Eliciting beta distributions for each π_i using the extended SHELF package

   For this step, unscaled (fixed lower bound of 0 and upper bound of 1) beta distributions were fitted. Based on prior experience with a similar antibiotic, for π_1, the expert gave a median value of 0.55, an upper quartile of 0.6 and a lower quartile of 0.5. The fitted beta distribution had parameters $Beta(d_1 = 25.4, e_1 = 20.8)$, a 10^{th} percentile of 0.46 and a 90^{th} percentile of 0.64. Upon being given feedback of the fitted beta distribution, the expert agreed that it was an agreeable representation of his original opinion about π_1. The expert gave wider bounds for π_2 as he/she was less confident about this quantity: a median of 0.3, an upper quartile of 0.35 and a lower quartile of 0.22. The fitted distribution was $Beta(d_2 = 6.51, e_2 = 15.5)$ with 10^{th} and 90^{th} percentiles of 0.18 and 0.42, respectively. For quantity π_3, the expert proposed a median of 0.15 with an upper quartile of 0.2 and a lower quartile of 0.11. The fitted beta distribution in this instance was $Be(d_3 = 4.46, e_3 = 23.6)$. The expert, again, accepted the fitted distributions for both π_2 and π_3 as reasonable representations of his/her knowledge.

3. Checking and adjusting the means

   The mean values summed up to 1.004 which is very close to 1 so no adjustment was necessary, i.e $d_i^* = d_i$, and $e_i^* = e_i$.

4. Finding the best choice of n

   Of the four possible values, $n_{med}, n_{mid}, \bar{n}$ and n_{opt}, $n_{opt} = 30.975$ was the chosen as the “best” by the facilitator. This choice implied more uncertainty about π_1 and less uncertainty about π_2 and π_3.

5. Providing feedback to the experts

   The expert decided to accept the choice of $Di(\mathbf{d}(n_{opt}))$ as an adequate representation of his/her knowledge since the original opinions were within the range of accuracy of that choice. The agreeable outcome of the elicitation was a Dirichlet distribution with $\mathbf{d} = (17.03, 9.162, 4.923)$.

Chapter 3

Data analysis, results and discussion

3.1 Practical application of the classical model

In this section, we explore how to apply performance-based weighting schemes for combining expert assessments using the classical model discussed in section 2.2.4.

We discussed 3 different types of weighting schemes: equal, global and item weighting. For equal weighting, we assign each density a weight of $\frac{1}{N}$ for N experts; for global weighting, the weights are determined per expert by the expert's calibration score and overall information score; and for item weighting, the weights are determined by the expert's calibration score, but these are calculated per expert and per variable. Once the weights are assigned, the actual parameters of interest are addresed by combining the experts' distributions for these parameters.

The following simple artificial example illustrates "global" weighting. "Equal" and "item" weighting would follow a very similar execution and are thus not illustrated below.

Suppose we have 3 experts: E_1, E_2 and E_3, and that each expert can give assessments which can be represented by a histogram with disjoint intervals of equal width. Consider the following summary data for E_1:

Interval	**Frequency (f_i)**	**Relative frequency**	**Mid-point (x_i)**	**Cumulative frequency (F_i)**
$[0-20)$	2	0.050	10	2
$[20-40)$	8	0.200	30	10
$[40-60)$	19	0.475	50	29
$[60-80)$	9	0.225	70	38
$[80-100)$	2	0.050	90	40
	Total $n = 40$	Total $= 1$		

Table 3.1: Histogram for expert E_1's assessments

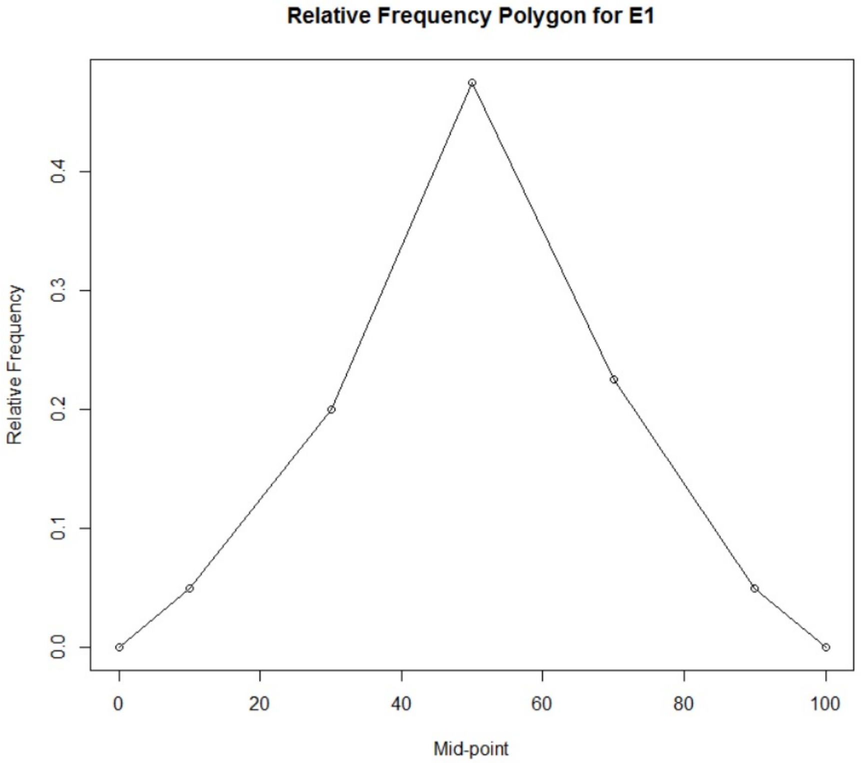


Figure 3.1: Relative Frequency Polygon for expert E_1

Parameter	Calculation
Mean	$\frac{1}{N}\sum f_i x_i = 50.5$
Variance	$\frac{1}{N}\sum (x_i - \bar{x})^2 = 100.03125$
Quartiles	$Q_k = L_i + \frac{\frac{k}{4}N - F_{i-1}}{f_i} a_i \text{ where} k = 1,2,3$ $\frac{k}{4}N = \frac{1}{4}40 = 10^{th} : Q_1 = 40$ $\frac{k}{4}N = \frac{2}{4}40 = 20^{th} : Q_2 = 50.5263$ $\frac{k}{4}N = \frac{3}{4}40 = 30^{th} : Q_3 = 62.2222$
Percentiles (5^{th},50^{th} and 95^{th})	$q_k = L_i + \frac{\frac{k}{100}N - F_{i-1}}{f_i} a_i \text{ where} k = 5,50,95$ $\frac{k}{100}N = \frac{5}{100}40 = 2^{nd} : q_5 = 20$ $\frac{k}{100}N = \frac{50}{100}40 = 20^{th} : q_{50} = 50.5263$ $\frac{k}{100}N = \frac{95}{100}40 = 38^{th} : q_{95} = 80$

Next, we calculated the relative information for E_1 as per equation 2.9:

p_i	interval	number of observed values in interval i	s_i
0.05	$0-20$	1	$s_1 = \frac{1}{12} = 0.0833$
0.45	$20-50.5263$	5	$s_2 = \frac{5}{12} = 0.4167$
0.45	$50.5263-80$	4	$s_3 = \frac{4}{12} = 0.3333$
0.05	$80-100$	2	$s_4 = \frac{2}{12} = 0.1667$
		Total = 12 seed values	

Therefore, the relative information, $I(s,p)$, is given by:

$$I(s,p) = \sum_{i=1}^{4} ln(s_i/p_i) = ln\left(\frac{0.0833}{0.05}\right) + ln\left(\frac{0.4167}{0.45}\right) + ln\left(\frac{0.3333}{0.45}\right) + ln\left(\frac{0.1667}{0.05}\right) = 1.3377$$

Calculating E_1's calibration score using equation 2.10 yields:

$$2NI(s,p) = 2(12)(1.3377) = 32.1048 \text{ and}$$
$$P(X^2_{(3)} > 32.1048) \approx 0.$$

Thus in a case like this, E_1 will not be used. We, however, continue with this example to illustrate how the rest of the process is applied.

Now to calculate the information, suppose the known value of the seed variable is $r = 25$. Also, suppose that a similar procedure to the one illustrated above was employed for experts E_2 and E_3, and their 5^{th} percentiles were reported as 23 and 25 respectively, while their 95^{th} percentiles were reported as 85 and 75 respectively. The lower and upper quantiles as seen in equation 2.11 are therefore given by

$$\begin{aligned} q_L &= min\{r, q_{5,1}, \ldots, q_{5,e}\} \\ &= min\{25, 20, 23, 25\} \\ &= 20, \end{aligned}$$

$$\begin{aligned} q_U &= max\{r, q_{95,1}, \ldots, q_{95,e}\} \\ &= max\{25, 80, 85, 75\} \\ &= 85. \end{aligned}$$

For expert E_1, supposing the analyst chose a $k\%$ over shoot value of $k = 10.8333$, we could calculate the 0^{th} and 100^{th} percentiles in equation 2.12 as follows:

$$\begin{aligned} q_0 &= q_5 - (k/100)(q_U - q_L) \\ &= 20 - (10.8333/100)(80 - 20) \\ &= 13.5, \end{aligned}$$

$$\begin{aligned} q_{100} &= q_{95} + (k/100)(q_U - q_L) \\ &= 80 + (10.8333/100)(80 - 20) \\ &= 86.5. \end{aligned}$$

Calculating expert E_1's information score (assuming a uniform distribution) as directed by equation 2.14 yields:

$$\begin{aligned} h_1 &= F_U(q_5) - F_U(q_0) = \frac{q_5 - q_0}{q_{100} - q_0} = \frac{20 - 13.5}{86.5 - 13.5} = 0.0890, \\ h_2 &= F_U(q_{50}) - F_U(q_5) = \frac{q_{50} - q_5}{q_{100} - q_0} = \frac{50.5263 - 20}{86.5 - 13.5} = 0.4182, \\ h_3 &= F_U(q_{95}) - F_U(q_{50}) = \frac{q_{95} - q_{50}}{q_{100} - q_0} = \frac{80 - 50.5263}{86.5 - 13.5} = 0.4037, \\ h_4 &= F_U(q_{100}) - F_U(q_{95}) = \frac{q_{100} - q_{95}}{q_{100} - q_0} = \frac{86.5 - 80}{86.5 - 13.5} = 0.0890, \end{aligned}$$

$$\begin{aligned} \therefore Inf = I(h,p) &= \sum_{i=1}^{4} p_i(p_i/h_i) \\ &= 0.05\left(\frac{0.05}{0.0890}\right) + 0.45\left(\frac{0.45}{0.4182}\right) + 0.45\left(\frac{0.45}{0.4037}\right) + 0.05\left(\frac{0.05}{0.0890}\right) \\ &= 1.0420\,. \end{aligned}$$

Below is a summary of the calculations for expert E_1 :

$$\begin{aligned} q_5 = 20 \quad &; \quad s_1 = 0.0833 \\ q_{50} = 50.5263 \quad &; \quad s_2 = 0.4167 \\ q_{95} = 80 \quad &; \quad s_3 = 0.3333 \\ &\quad\ \ \, s_4 = 0.1667 \end{aligned}$$

$$\begin{aligned} I(s,p) &= 1.3631 \\ 2NI(s,p) &= 32.1048 \\ Cal &= P(X^2_{(3)} > 32.1048) \approx 0 \\ Inf &= 1.042\,. \end{aligned}$$

Similar summaries can be shown for experts E_2 and E_3 using artificial estimates of the quantiles q_5 and q_{95}, and the s_i's. For expert E_2 :

$$\begin{aligned} q_5 = 23 \quad &; \quad s_1 = 0.042 \\ q_{50} = 50.5263 \quad &; \quad s_2 = 0.458 \\ q_{95} = 85 \quad &; \quad s_3 = 0.44 \\ q_0 = 16.5 \quad &; \quad s_4 = 0.06 \\ q_{100} = 91.5 & \end{aligned}$$

$$\begin{aligned} I(s,p) &= ln\left(\frac{0.042}{0.05}\right) + ln\left(\frac{0.458}{0.45}\right) + ln\left(\frac{0.44}{0.45}\right) + ln\left(\frac{0.06}{0.05}\right) = 0.0031 \\ 2NI(s,p) &= 24(0.0031) = 0.0744 \\ Cal &= P(X^2_{(3)} > 0.0744) = 0.9947 \text{ (a high calibration score)} \\ Inf &= 0.05\left(\frac{0.05}{\frac{23-16.5}{91.5-16.5}}\right) + 0.45\left(\frac{0.45}{\frac{50.5263-23}{91.5-16.5}}\right) + 0.45\left(\frac{0.45}{\frac{85-50.5263}{91.5-16.5}}\right) + 0.05\left(\frac{0.05}{\frac{91.5-85}{91.5-16.5}}\right) = 1.05; \end{aligned}$$

and for expert E_3 :

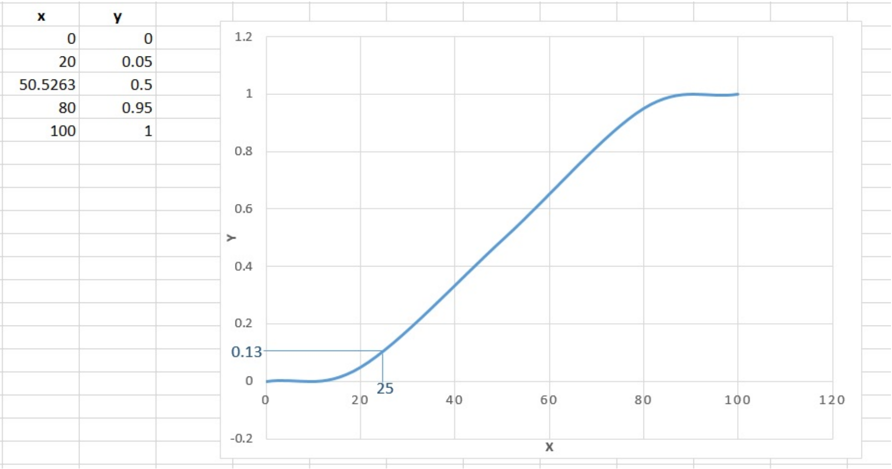
$$\begin{aligned} q_5 = 25 \quad &; \quad s_1 = 0.07 \\ q_{50} = 50.5263 \quad &; \quad s_2 = 0.43 \\ q_{95} = 75 \quad &; \quad s_3 = 0.46 \\ q_0 = 18.5 \quad &; \quad s_4 = 0.04 \\ q_{100} = 81.5 \quad & \end{aligned}$$

$$\begin{aligned} I(s,p) &= ln\left(\frac{0.07}{0.05}\right) + ln\left(\frac{0.43}{0.45}\right) + ln\left(\frac{0.46}{0.45}\right) + ln\left(\frac{0.04}{0.05}\right) = 0.0898 \\ 2NI(s,p) &= 24(0.0898) = 2.1552 \\ Cal &= P(X^2_{(3)} > 2.1552) = 0.5406 \,\text{(an average calibration score)} \\ Inf &= 0.05\left(\frac{0.05}{\frac{25-18.5}{81.5-18.5}}\right) + 0.45\left(\frac{0.45}{\frac{50.5263-25}{81.5-18.5}}\right) + 0.45\left(\frac{0.45}{\frac{75-50.5263}{81.5-18.5}}\right) + 0.05\left(\frac{0.05}{\frac{81.5-75}{81.5-18.5}}\right) = 1.07. \end{aligned}$$

Now we can calculate the weights using $\alpha = 0.05$; and letting $1_\alpha = 0$ if the calibration score is less than α, and 1 otherwise:

$$\begin{aligned} w_\alpha(e) &= 1_\alpha(calibration\,score) \times calibration\,score(e) \times information\,score(e) \\ \therefore \text{For}\, E_1 : \quad w_\alpha(e) &= 0(0)(1.042) = 0 \\ \text{For}\, E_2 : \quad w_\alpha(e) &= 1(0.9947)(1.05) = 1.0444 \\ \text{For}\, E_3 : \quad w_\alpha(e) &= 1(0.5406)(1.07) = 0.5784\,. \end{aligned}$$

To finally calculate the decision maker score, we first to need to calculate $f_{e,i}$ as per equation 2.8 for each expert. We show how to obtain this value for E_1, and artificial values are assigned to E_2 and E_3.



x	y
0	0
20	0.05
50.5263	0.5
80	0.95
100	1

Figure 3.2: Expert E_1's density function

Figure 3.2 shows that $f_{1,i} = 0.13$. Assuming $f_{2,i} = 0.26$ and $f_{3,i} = 0.19$, we can calculate the DM score using equation 2.16 as follows:

$$\begin{aligned} DM_\alpha(i) &= \frac{\sum\limits_{e=1,\ldots,E} w_\alpha(e) f_{e.i}}{\sum\limits_{e=1,\ldots,E} w_\alpha(e)} \\ &= \frac{0(0.13)+1.0444(0.26)+0.5784(0.19)}{1.0444+0.5784} \\ &= 0.2351\,. \end{aligned}$$

Therefore, 0.2351 is the global weight combined DM score where our choice of α maximises the product of the calibration and the information score.

Chapter 4

Conclusion

The traditional committee method and the Delphi method are both “census” methods of expert elicitation. They employ behavioural aggregation methods and, as a result, are prone to psychological bias. For this reason, they are no longer popular methods of expert judgment analysis.

Paired comparison methods such as the Negative Exponential Life (NEL) model are more popular as they are quite thorough in ensuring that all possible pairs are compared and ranked, and the comparisons can be further reduced into an interval or ratio scale. In section 2.2.3, the use of an illustration of how the NEL model can be applied enabled us to develop formulas that can be used to represent “no agreement” and “total agreement”, which can be found in equations 2.2 and 2.3, respectively. Furthermore, we were able to relate the coefficients of concordance (w and w') for a small and large number of items (n) being compared among some experts (e) using $w = kw'$, where $k = \frac{1}{e(n-1)}$. The extensive comparing done in this method, however, opens it up to issues of redundancy, and as with the traditional committee method and the Delphi method, there is no assessment of uncertainty that is carried out.

Cooke’s classical model results in “rational consensus” since the experts are judged on their ability to judge uncertainties, and this is what makes it the most widely used technique to date. It employs mathematical aggregation techniques to combine expert opinion and does not aim to impose agreement but rather to quantify and combine opinions on uncertainty. The illustration of this model performed in chapter 3 showed that it is a method that is carried out step by step, thus allowing disparities of opinion and loopholes in the experts’ knowledge to be discovered early. The experts’ calibration/accurateness and informativeness are weighted and combined to ultimately produce a Decision Maker (DM) score of the uncertainty regarding the unknown variable. Cooke’s method is regarded as “classical” as it is a frequentist approach to expert judgment analysis.

In the increasingly popular Bayesian approach, after giving their opinions about the unknown variable, experts are further allowed to assess other experts' opinions and adjust their estimates. The first revision of estimates is said to produce "improved estimators", and any further revisions produce mere "compromise estimators", as shown by the illustration performed in section 2.2.5. In this approach, consensus refers to the revision with the smallest distance between the maximum and minimum estimates of a parameter. During this process, it is unlikely that experts will achieve the same estimate, rather, each expert's estimate will converge to an unbiased variation of the other experts' original opinions, as illustrated by the example in section 2.2.5.

The Dirichlet distribution is becoming a common tool in multivariate elicitation tasks where the facilitator needs information about a set of proportions that must sum up to 1. This distribution, unfortunately, only has one parameter (n, the sum of the uncertain quantities) that controls uncertainty and also measures the experts' informativeness. A technique called overfitting is used to elicit more judgments than needed from the expert in order to better identify an appropriate choice of $\mathbf{d}$, the parameter vector of uncertain quantities. In maintenance optimisation, the histogram technique is used to display the distribution of the probabilities of failure in fixed time intervals as it is easier than trying to fit an entire parametric life distribution. In multinomial cases, the Supra Bayes approach is then used to combine the likelihood function for the expert judgments while incorporating their background information, with a prior distribution for the quantity of interest, to ultimately obtain a Generalised Dirichlet distribution as a posterior distribution. It is also now acceptable to not necessarily aim to fit the Dirichlet distribution to an expert's beliefs, and to reach a final resolve that no Dirichlet distribution was an acceptable representation of an expert's opinion.

Areas of further research

When uncertainty is quantified, it is always conditional on something that can be elicited from its background information. Cooke and Goosens (2010) write that a "case structure" is in place during the elicitation process to ensure that this background information is assessed in a way that does not introduce too much "noise" into the process from the various different backgrounds of the experts. Since it is difficult to exhaust the list of all possible relevant variables, uncertainties due to any unknown values of unspecified variables are "folded into" the uncertainties of the target variables. Cooke and Goosens (2010) warn that this can cause dependencies in the target variables' uncertainties and dealing with dependence is an area still under development. For now, experts are asked to identify groups of variables that may have significant dependence when specifying their subjective distributions.

There also needs to be more research towards the appropriate construction and use of seed questions. Further research could explore constructing seed questions carefully such that they weed out some of the structural bias present in expert selection as discussed in section 2.2.4.

Zapata-Vazqeuz et al. (2012) suggest that more could be explored in the area of multivariate elicitation of a joint distribution for a set of proportions, besides using the Dirichlet distribution as it does not always prove to be adequate.

References

Aspinall, W. (2010). A route to more tractable expert advice. *Nature 463*, 294–295.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Cooke, R. M. and L. H. J. Goosens (2008). TU delft expert judgment data base. *Reliability Engineering and System Safety 93*, 657–674.

Cooke, R. M. and L. H. J. Goosens (2010). Expert judgment elicitation for risk assessments of critical infrastructures. *Journal of Risk Research 7*, 643–656.

de Franca Doria, M., E. Boyd, E. L. Tompkins, and W. N. Adger (2009). Using expert elicitation to define successful adaptation to climate change. *Environmental Science and Policy 12*, 810–819.

Eggstaff, J. W., T. A. Mazzuchi, and S. Sarkani (2014). The effect of the number of seed variables on the performance of cooke's classical model. *Reliability Engineering and System Safety 121*, 72–82.

EMSE280 (2011, Spring). Expert judgment - techniques of risk analysis and management. The George Washington University.

Garisch, I. (1985). *Multi-Bayesian decisions under a Super - Bayesian Approach*. PhD. Thesis.

Mayer-Schonberger, V. and K. Cukier (2013). *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mazzuchi, T. A., W. G. Linzey, and A. Bruning (2008). A paired comparison experiment for gathering expert judgment for an aircraft wiring risk assessment. *Reliability Engineering and System Safety 93*, 722–731.

Ryan, J. J. C. H., T. A. Mazzuchi, D. J. Ryan, J. Lopez de la Cruz, and R. Cooke (2012). Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research 39*, 774–784.

Siegel, S. (1956). *Nonparametric Statistics*. New York, NY.

Van Noortwijk, J. M., R. Dekker, R. M. Cooke, and T. A. Mazzuchi (1992, September). Expert judgment in maintenance optimization. *IEEE Transactions on Reliability 41*(3), 427–432.

Walker, K., J. Neumann, H. Roman, and T. Gettleman (2004, November). Appropriate number of experts for the PM EJ project.

Zapata-Vazqeuz, R. E., A. O'Hagan, and L. S. Bastos (2012, December). Eliciting expert judgments about a set of proportions. *Journal of Applied Statistics 00*, 1–15.