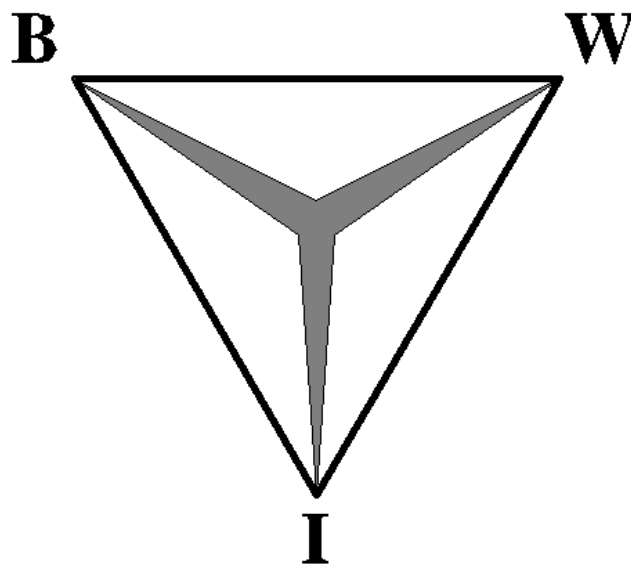


Expert opinion

Use in practice

BWI-werkstuk



Ivo Roest
Vrije Universiteit Amsterdam
October 2002

Contents

Preface	3
Abstract	4
Chapter 1 Introduction	5
§1.1 About knowledge, uncertainty and ignorance	5
§1.2 What is uncertainty?	7
§1.3 What is expert opinion?	7
H2 History of expert opinion: An overview	10
§2.1 How it all began	10
§2.2 Scenario analysis	10
§2.3 Delphi Method	11
§2.4 Cross impact analysis	12
H3 Use in practice	13
§3.1 Characteristics/(Dis)Advantages of expert opinion	13
§3.2 Expert opinion as "data"	14
§3.3 Structured methods	15
§3.4 The structured process of elicitation	15
§3.5 Annotations on input process	17
§3.6 Elicitation	18
§3.7 Scoring rules and weighting of opinions	18
H4 Models & mathematical backgrounds	20
§4.1 Entropy, (relative) information and calibration	20
§4.2 Scoring rules	21
H5 Models to combine expert's opinions	22
§5.1 Bayesian models	22
§5.2 Physical scaling models	22
§5.3 Classical model	23
H6 Conclusions	26
Bibliography	27

Preface

During the last year of the study Business Mathematics and Computer Science the students perform a literature search. This study can be at any subject. The subject of this study is expert opinion (or expert judgement/expert judgment).

Where did I get this subject from? To explain, go back in history:

In the year 431 B.C. the Greek general Pericle spoke the next few words to his soldiers, before starting war against the Spartans^[1]: "(...) The worst thing is to rush into action before the consequences have been properly debated. (...) We are capable at the same time of taking risks and estimating them beforehand. Others are brave out of ignorance, and when they stop to think, they begin to fear. But the man who can most truly be accounted brave is he who best knows the meaning of what is sweet in life, and what is terrible, and he then goes out undeterred to meet what is to come."

This is what risk assessment and management is about: assess possible future events (risks), assigning weights to them, estimate their effect. During my apprenticeship with Deloitte & Touche Enterprise Risk Services, my coach Coby Peeters-Vergeer worked on an assignment for a big government agency. To support her with the technical aspects of this assignment, I chose expert opinion as the subject of my literature study.

In this paper, I will try to answer the following questions:

1. What is uncertainty; how do people reason with it?
2. What is expert opinion?
3. What are the basic concepts of expert opinion?
4. Which mathematical concepts "play a part" in expert opinion?

This paper intends to give a comprehensive, but to-the-point overview of the field of expert opinion.

I would like to say a few words of gratitude to a few people, who helped me compiling this paper by giving useful comments and feedback: my coach Coby Peeters-Vergeer from Deloitte & Touche ERS, and my coach from the Vrije Universiteit, Geurt Jongbloed.

I hope you will enjoy reading this paper. I enjoyed working on it, because the subject itself was completely new to me.

Amsterdam, October 2002.

Ivo Roest
iroest@deloitte.nl

Abstract

Expert opinion plays a major role in assessing problems for which data are lacking. Experts may have valuable knowledge about problems and solutions in their field. Although this knowledge is not certain (it is entertained with a level of subjective confidence); elicitation, quantification and aggregation of these expert opinions may provide important knowledge to decision makers in many fields of politics, science and technology. In practice, often this is done in an unstructured and ad hoc way.

Literature stresses that it is important, when conducting an expert opinion elicitation process, to do this in a structured, clear and transparent way. The principles fairness, neutrality, accountability and empirical control should be taken care of. Formal procedures that obey these principles, provide many possibilities and advantages to decision makers.

Experts may have valuable knowledge about problems in their field. Although this knowledge is not certain (it is entertained with a level of subjective confidence); elicitation, quantification and aggregation of these expert opinions may provide important knowledge to decision makers in many fields of politics, science and technology.

Due to better defined (and structured) methods of elicitation and an improvement of the mathematical foundations for the use of expert opinion, its use in practice at the national and international level is growing. After a period of relative silence in the seventies and eighties, the elicitation and analysis of expert opinion nowadays plays an increasingly important role in issues at different levels of decision making.

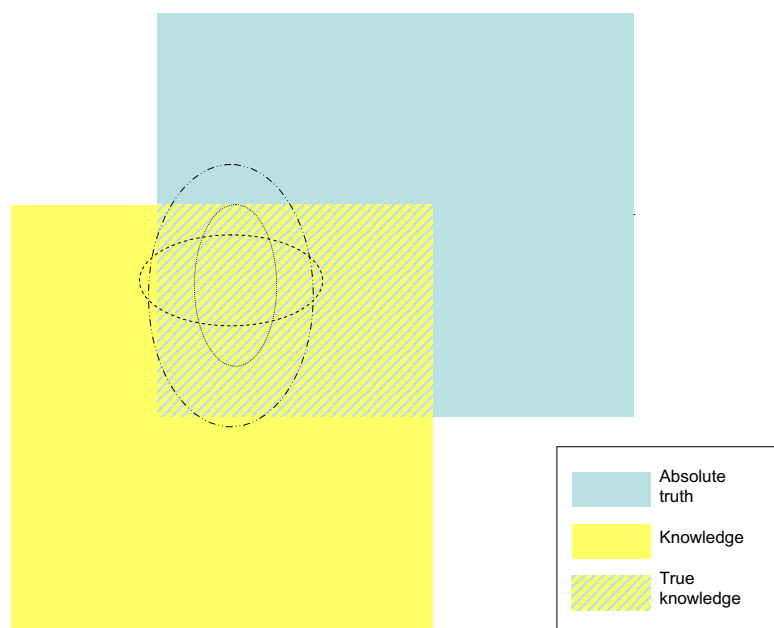
Chapter 1 Introduction

§1.1 About knowledge, uncertainty and ignorance

Ayyub^[7] describes **knowledge** as the body of truth, information and principles acquired by humankind about some subject of interest. **Information** is a subset of knowledge, that is acquired by investigation and research. These previous two terms might not describe the whole state of the subject of interest (the absolute truth). Knowledge will always reflect the imperfect nature of humankind, that can be attributed to:

- their reliance on senses for knowledge acquisition;
- their mind for extrapolation, creativity and imagination; and
- their preconceived opinion (due to time asymmetry: our limited capacity to free ourselves from the past in order to forecast in the future and our inability to go back in time and verify historical claims, therefore it gives us overconfidence in the superiority of our present knowledge).

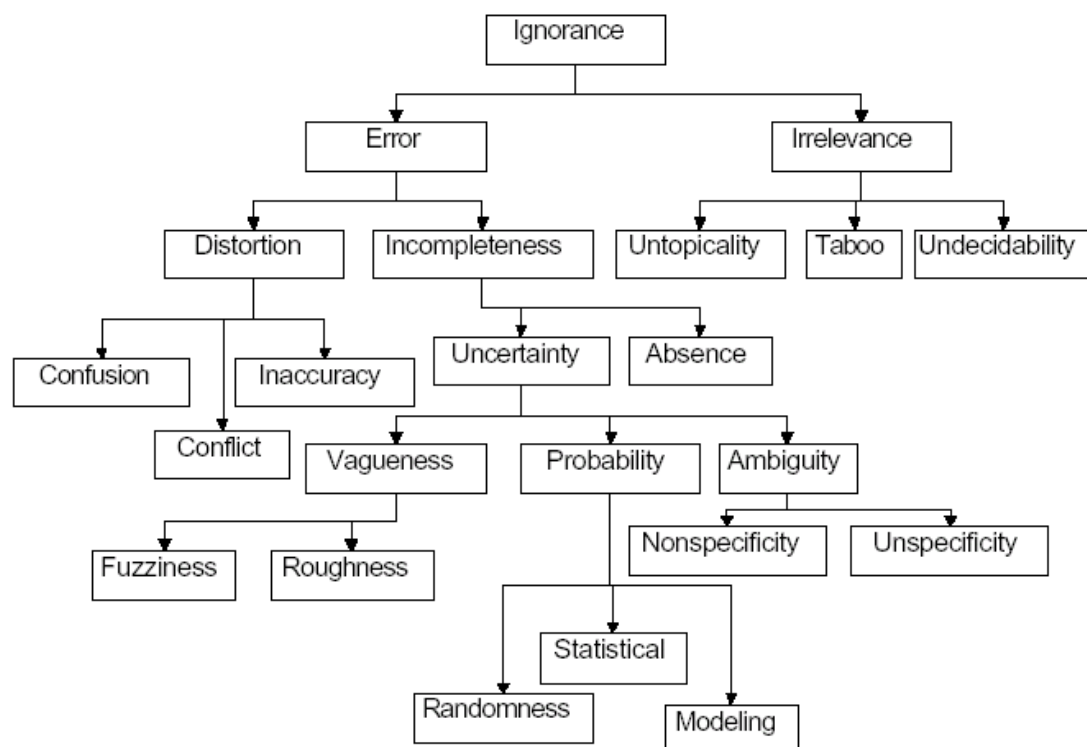
Knowledge is primarily the product of the past as we know more about the past than the future. Our inability to go back in time gives most of us overconfidence in the superiority of our present knowledge. Furthermore, humans tend to focus on what is known, and brush aside ignorance. As knowledge is a mixture of truth and fallacy^[7], there exist two types of ignorance: within the knowledge base, and outside the knowledge base. This can be represented as follows:



Figuur 1: Knowledge vs. ignorance

In the figure above there are some ellipses. An ellipse can represent the true knowledge of an expert, the self-perceived knowledge of an expert, and the perception by others of the expert's knowledge. If the smallest ellipse represents the true knowledge of an expert, and the biggest ellipse the self-perceived knowledge, then the difference between the two of them is the **overconfidence** of the expert.

Ayubb^[7] divides ignorance into two main factors: error (being ignorant of something (unintended)) and irrelevance (ignore something deliberately). This can be represented as follows:



Figuur 2: Ignorance specified

Cooke^[2] also mentions the term **error**: he thinks people do not usually perform mental calculations, but rely instead on various rules of thumb (heuristics). By error he means a violation of the basic axioms of probability, or an estimate that is not really in accord with an expert's beliefs. When heuristics lead to errors in this sense, he speaks of **biases**. (Beware! Sometimes the term bias is used to refer to the willfully distortion of expert opinion.)

Uncertainty can be further classified. This is done in the next paragraph.

§1.2 What is uncertainty?

According to Hogarth^[6], uncertainty is a description of the imperfect knowledge of the true value of a particular variable or its real variability in an individual or a group. In general, uncertainty is reducible by additional information-gathering or further analysis.

Uncertainty can be classified into four different types:

1. **Variable uncertainty:** This occurs when variables used in the analysis, cannot be measured accurately/precisely because of, for example, equipment limitations or temporal variances between the measured quantities.
2. **Model uncertainty:** This kind of uncertainty is related to all models used in risk analysis. For example, computer models used to predict a certain quantity, are simplifications of reality. They exclude some variables that influence predictions, but are hard to include because of increased complexity of the model, or a lack of data for these variables.
3. **Decision-rule uncertainty:** This type of uncertainty arises out of the need to balance different social concerns when determining an acceptable level of risk. Uncertainty concerning risk analysis influences many risk management decisions. What are the possible outcomes of a decision?
4. **Variability:** This is associated with the variations in physical and biological processes, and cannot be reduced by additional information- gathering or further analysis.

Managing uncertainty, and making decisions requires knowledge. This knowledge can eventually be gathered by expert opinion.

§1.3 What is expert opinion?

Risk analysis and making decisions on the basis of this requires the knowledge of two main quantities for components/systems/etcetera: the **probability of occurrence** of a future event, and the **size of its consequences**.

According to Stamatelatos^[1], risk is defined by the following scheme (for rare events, frequency is the probability of occurrence per unit time):

$$\begin{array}{|c|} \hline \mathbf{RISK} \\ \hline \frac{\text{Detriment}}{\text{Unit Time}} \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{FREQUENCY} \\ \hline \frac{\text{Events}}{\text{Unit Time}} \\ \hline \end{array} \times \begin{array}{|c|} \hline \mathbf{SEVERITY} \\ \hline \frac{\text{Detriment}}{\text{Event}} \\ \hline \end{array}$$

Figuur 3: Mathematical expression of risk

But how to obtain these two quantities? There are multiple possibilities to do this; for example with the aid of statistics. One other possible solution is to use expert opinion.

Expert opinion is the judgement, based on knowledge and experience, that an expert makes in responding to certain questions about a subject. These questions can be related to probabilities, ratings, uncertainty estimates, weighting factors, physical quantities, enz.

An **expert** is a key person who:

- has important knowledge about the field of interest;
- has a background in the field of interest;
- is recognized (such as by his colleagues) as qualified to address problems in the subject area; and
- has familiarity with probability assessments (not at any price: this can be given by training).

Expert opinion can be viewed as a representation of an expert's state of knowledge at the time of response to the technical question. Thus, expert opinion should change through time as the expert receives new information.

Expert opinion is used in all kinds of technical fields - medicine, economics, engineering, risk/safety assessment, knowledge acquisition, decision sciences, pharmaceuticals, environmental studies, et cetera.

Some examples for which expert opinion was used: a doctor must determine the likelihood of a patient's illness based on the advice of four different specialists; the executives of an investment company must decide which of several stocks to purchase at a given time.

Expert opinion is used in two ways:

1. To structure a problem. Experts determine which data and variables are relevant for analysis, which analytical methods are appropriate and which assumptions are valid. Statisticians frequently use their expert judgment in this way.
2. To provide estimates. For example, experts may estimate failure or incidence rates, determine weighting factors for combining data sources, or characterize uncertainty.

Due to better defined (and structured) methods of elicitation and an improvement of the mathematical foundations of expert opinion, its use in practice at the national and international level is growing. After a period of relative silence in the seventies and eighties, the elicitation and analysis of expert opinion nowadays plays an increasingly important role in issues at different levels of decision making.

This paper provides in the next chapters methods for eliciting expert opinion on possible future events, their probabilities and their consequences. Historical background on the development of expert opinion elicitation is provided in chapter 2.

Its limitations, current uses, disadvantages/advantages of using and guidelines how to use expert opinion, are discussed in chapter 3.

The mathematical backgrounds are provided in chapter 4, and some models in Chapter 5.

H2 History of expert opinion: An overview

§2.1 How it all began

The use of experts in decision making is not completely new. But, the use of expert opinion in a structured way is relatively new.

After World War II a period occurred of rapid growth in the field of Research and Development (R&D). During this period there was some kind of "honeymoon"^[2] between the United States government and several US universities and institutions. The US government spent much money into getting advice (from think tanks) on a broad range of subjects: strategic and tactical planning, sociology, international relations, new technologies, et cetera.

The close relationship between government and science ended at the beginning of the seventies, during the Vietnam War. Most scientists were opposed to the war, and resigned from advisory bodies.

The foundation of expert opinion in those days was laid by a company called RAND Corporation. This company developed two very important methodologies for elicitation of expert opinion: the Delphi method and scenario analysis. RAND had as an exponent Herman Kahn. The latter is regarded as the father of **scenario analysis**.

§2.2 Scenario analysis

Herman Kahn^[3] is regarded as the founder of scenario analysis. A scenario is a possible sequence of events. Scenario analysis consists of the next few steps:

1. The analyst identifies, what he thinks is, a set of basic long-term trends.
2. These trends are then extrapolated into the future, taking into account any knowledge that might have impact on such extrapolations. The result of this step is called the **surprise-free scenario**.
3. Based on this, other **alternative scenarios** can be defined, varying key parameters in the surprise-free scenario.

Scenario analysis does not yield any probabilities. All the scenarios have the same probability to occur. One could wonder what could be the use of studying the surprise-free scenario, and a few alternatives, if an analysis of these doesn't yield any probabilities. Studying the surprise-free scenario could help getting better comprehension of the basic trends.

§2.3 Delphi Method

The **Delphi method**^[4] was developed during the early 1950s, and was the first structured method for eliciting and combining expert opinion. The Delphi method is based on a structured process for collecting and distilling knowledge from a group of experts by means of a series of questionnaires combined with controlled opinion feedback

A group of experts are asked individually to provide their views on what will happen in the future.

The process:

1. Each expert gives his independent opinion on a list of questions.
2. The opinions of each expert are collated. Extreme opinions are discarded, and an initial view (consensus) is formulated.
3. The initial view is circulated to the experts for their further comments, and depending on how they respond, the initial view might be changed.
4. The process will continue until a prediction for the future has been made, which has the acceptance of all/most of the panel of experts.

Makridakis and Wheelright^[4] summarize general complaints against the Delphi method:

1. A low level reliability of opinions among experts and therefore dependency of forecasts on the particular judges selected.
2. The sensitivity of results to ambiguity in the questionnaire that is used for data collection in each round.
3. The difficulty in assessing the degree of expertise incorporated into the forecast.
4. Experts tend to judge the future of events in isolation from other developments.
5. The responses can be altered by the monitors in the hope of moving the next round responses in a desired direction.

There have been many poorly conducted Delphi projects. However, there is a big difference between evaluating a technique and evaluating an application of a technique. There have been several studies. A study by Basu & Schroeder^[4] compared Delphi forecasts of five-year sales with both unstructured, subjective forecasts and quantitative forecasts that used regression analyses. When compared with actual sales for the first two years, errors of 3-4% were reported for Delphi, 10-15% for the quantitative methods, and of approximately 20% for the previously used unstructured, subjective forecasts.

Gordon^[9] notes, that an improvement in forecasting reliability over the Delphi method was thought to be made by taking into consideration the possibility that the occurrence of one event may cause an increase/decrease in the probability of occurrence of other events included in the survey. Therefore **cross impact analysis** has been developed, as an extension of Delphi techniques.

§2.4 Cross impact analysis

As described in the preceding paragraph, a basic limitation of the Delphi method, and many forecasting techniques, is that they only give separated forecasts, i.e. events and trends are considered separately from each other; without thinking of their possible influence on each other.

Cross impact analysis is developed by Gordon & Helmer in 1968. In the beginning, it was meant as a concept for a forecasting game.

Cross impact analysis is a stepwise process, which consists of the following steps:

1. Define the events to be included in the study.
2. Define the time planning interval.
3. Develop matrices to define the interdependencies between events and trends.
4. Estimate the initial probability of each event.
5. Perform a calibration run.
6. Define tests and actions to be run with the matrix.
7. Perform calculations.
8. Evaluate results.

The calculations are performed repeatedly, until the probabilities converge to some value (the experts agree on).

H3 Use in practice

§3.1 Characteristics/(Dis)Advantages of expert opinion

Expert opinion is typically elicited and analyzed when:

1. Data are sparse or difficult to obtain. Sometimes information is not available from historical records, prediction methods or literature. This can occur when activities of an enterprise create new conditions and circumstances, without useful data for analysis. Expert opinion may be used to provide estimates on these new, rare, complex, or poorly understood problems.
2. Data are too costly to obtain.
3. Data are open to different interpretations. Results are unstable/uncertain.
4. Models to analyse risks are not available; or are very data intensive.
5. There is a need to perform an initial screening of problems^[7]. It is used to determine the state of knowledge in a problem (i.e., what is known and how well it is known) and document that information, such as in a data or knowledge base.

There are some (dis)advantages of using expert opinion:

Advantages:

1. It can be a low-cost method.
2. Quick method.
3. Relies on knowledgeable, experienced people.

Disadvantages:

1. One or more members may dominate the group.
2. Experts may be incompetent.

Expert opinion can be expressed in two forms:

- **Quantitative** form: This can be probabilities, ratings, uncertainty estimates, weighting factors, and physical quantities (e.g., costs, time, length, weight, etc.).
- **Qualitative** form: A textual description of the expert's assumptions in reaching an estimate, reasons for selecting or eliminating certain data or information from analysis, and natural language statements of physical quantities of interest.

§3.2 Expert opinion as "data"

When expert opinion is in quantitative form, it can be considered to be "data". Expert opinion has some characteristics in common with empirical data from experiments or physical observations^[2]:

- Expert opinion is affected by the process of gathering it. Elicitation methods take advantage of the body of knowledge on human cognition, and include procedures for aiding memory and countering effects arising from the phrasing of the questions, response modes, the influence of the elicitor, and the expert's personal agenda, the information the experts considered, the experts' methods of solving the problem, and the experts' assumptions. Choosing the wrong method may lead to bad results.
- Expert opinion has uncertainty, which can be characterized and subsequently analyzed. Many experts are accustomed to giving uncertainty estimates in the form of simple ranges of values. In eliciting uncertainties, the analysts can make experts aware of their natural tendency to underestimate uncertainty, such as through the exercise of estimating on sample problems.

The main difference between expert opinion and empirical data is that expert opinion is a form of personal opinion (e.g. **subjective probabilities**). Cooke^[2] points out there is a difference between the subjective and "normal probabilistic" view on probabilities; but that this doesn't mean the probabilistic view is more objective than the subjective view. The proof for this is delivered by Savage's normative decision theory and De Finetti's theory of probability^[2]. (Note: one should be very careful in treating subjective probabilities as if they're just normal probabilities; because they ain't!)

When expert opinions are being used as data, Cooke^[2] sees some trends:

1. Expert opinion estimates typically show a wide spread.
2. Estimates, given by the experts, are not always independent. For example: if an expert judges negative of one aspect within a study, then he could also have a tendency to be negative of other aspects within the study, too.
3. In common, if the same expert opinion methodology is applied several times on the same problem, it doesn't produce similar results.
4. Mostly, the subjective probabilities don't agree at all with observed frequencies.

He thinks the trends mentioned above are due to the fact that expert opinion has been used unstructured; without the use of formal processes/methodologies, and emphasizes that the use of structured methods can deliver many advantages.

§3.3 Structured methods

What is the goal of applying structured expert opinion methods? The main objective is to enhance rational consensus. Goossens & Cooke^[8] see a few conditions to achieve this objective:

- **Accountability:** All used data and applications must be open to peer review by competent independent reviewers. The produced results must be reproducible.
- **Empirical control:** Estimates provided by experts are compared to empirical control data.
- **Neutrality:** The method, used during an expert opinion elicitation project, should encourage experts to state their true opinions, and not bias the results.
- **Fairness:** The experts are not pre-judged.

These conditions have been applied in certain models, e.g. the classical model. These models are treated in Chapter 4.

But what does an elicitation process look like?

§3.4 The structured process of elicitation

According to Ayyub^[7], expert opinion elicitation is a formal, heuristic process of obtaining information or answers to specific questions about certain quantities (such as expected service life of a product) and probabilities of future events.

Expert opinion elicitation techniques are techniques which involve interviewing experts, and asking them to assess unknown quantities, or probabilities of possible future events. In a panel a group of experts can discuss the future, and make forecasts.

Rantilla & Budescu^[5] describe two approaches to aggregation of expert opinion:

1. **Normative approach:** Normative models are predictive and prescriptive. Based on a set of assumptions and rules, these models suggest what to do in a given situation to optimize a well-defined objective function. Solutions of these models are evaluated by using actual responses from practical experiments.
2. **Descriptive approach:** In this perspective, the basic assumption is that it is possible to find some common process underlying decision making, by constructing models from empirical data.

The structured process of expert opinion elicitation can be conducted in various ways, but in this chapter I will give a description, made by Cooke^[12].

In paragraph 3.5 and further some additions will be made on the steps to be taken during the process.

- Preparation of the expert elicitation process:
 1. Definition of the case structure document: This document contains the description of the field of interest, what is expected from the experts, and in what way which experts will be queried about which problems.

The document contains the frame for the panel of experts, specifying all issues to take into account, while conducting the expert elicitation process.
 2. Identification of variables:
 - **Target variables:** Which variables are to be quantified through by the experts?
 - Identification of the **query variables:** These are the variables to be assessed by the experts. The target variables may not be appropriate for direct elicitation. Then it is needed to do find derived variables for these.
 - Identification of the **seed variables:** Variables whose true values are unknown to experts when giving their opinion, but whose values are known post hoc.
 3. Identification and selection of experts: Choosing experts, and selecting from the initial list of experts the final group for the elicitation process, on the basis of a few selection criteria: reputation in the field of interest, diversity in background, familiarity with uncertainty concepts.
 4. Defining the exact questions and format for elicitation.
 5. Test run to try out the questionnaire and its format.
 6. Training/preparation of the experts for the probabilistic assessments: The experts will provide assessments over the query variables in terms of quantiles, for instance 5%, 50%, 95%. Most experts are not familiar with stating their opinion over variables.
- Elicitation of expert opinions:
 7. A session in which each expert is interviewed individually, or a joint meeting where the experts' individual opinions (delivered earlier) are discussed in the presence of a few analysts, which are experienced in probability issues, and the field of interest.

- Handling the results of the elicitation session:
 8. Scoring the expert opinions/combining the assessments: weighting the experts by using one of the methods described later. This can be done by software packages. One of these packages is Excalibr, made by researchers of the university of Delft.
 9. Robustness analysis of the combined results. This is done by removing experts/seed variables from the dataset one at the time, and recalculating the decision maker. If the relative information loss with respect to the original decision maker is large, the results may not be replicated if another study were to be done using other expert and seed variables.
 10. Discrepancy analysis of the combined results. If there are items in the study on which the opinions of the experts differ most, these items should be reviewed.
 11. Translating back the uncertainties of the combined expert assessments on the query variables into uncertainties on the target variables from step 2.
 12. Documentation of the results and feedback with the experts.

§3.5 Annotations on input process

The next few characteristics of the input have an important impact on the final results:

1. Characteristics of the experts: e.g. reliability, accuracy, expertise, experience, background. When making decisions about weighting the opinions of different experts, decision makers pay attention to the accuracy levels of the experts, and the confidence levels the experts give for their own opinions.
2. Redundancy in the information the experts use: Hogarth^[6] points out that decision makers should be more confident if they combine information from multiple sources that are not redundant and highly credible. Confidence should drop, when redundancy or credibility are low.
3. Inter-correlation between various opinions of the experts: Usually, it is desirable to reduce the correlation between experts, because this reduces the overall accuracy. Research^[6] showed that adding more experts doesn't reduce the overall accuracy. Moreover, Rantilla & Budescu^[5] showed decision makers were more confident when there were fewer experts. This may be due to the fact when experts were added, the amount of disagreements increased.
4. Amount of information available.

§3.6 Elicitation

Elicitation can be done in three ways:

1. **Direct elicitation:** This is the simplest method of measuring a person's degree of belief. You just ask him what his degree of belief is. This method is the most common, but it is also the worst to use. Most experts don't know how to handle probabilities. For example, the experts are asked to give estimations of:

- **Median** value M : this is a measure for central tendency. It is defined as the point that divides the data in two equal parts: 50 percent of the data are below this point, and 50 percent are above this point:

$$M = x_{((n+1)/2)}$$

Sometimes the mean value is used. But one can prefer the use of the median value, because this is insensitive to extreme values.

- **Percentiles:** often this are the 5 percent percentile and the 95 percent percentile of the data. A **percentile** x_q is the value of a random variable such that q percent of the data is less or equal to x_q .
2. **Indirect elicitation:** Instead of asking people their opinion in the form of probabilities, you ask them questions on which the answers can be translated into probabilities.

§3.7 Scoring rules and weighting of opinions

Basically, three forms exist for aggregating expert opinions: by a model, by one decision maker, by a group of decision makers. But one can make various combinations of these three basic forms^[5].

1. **By a model:** A series of expert opinions function as inputs to a (normative) model. The model generates a solution to the problem.
2. **By one decision maker:** Based on the expert opinions offered as inputs, a single decision maker produces an output (descriptive model).
3. **By a group of decision makers:** A panel of experts first determines the inputs, and afterwards discuss these together, until some form of consensus has been reached. This is the output of the process.

The most common approaches to aggregation are:

- Use of one simple decision maker, using inputs from experiments.
- Use of one researcher, who uses data provided by various panels of experts.

Scoring rules are rules to assess the information reliability and quality provided by experts through an expert-opinion elicitation process. The scores from these rules can be used to determine weight factors for combining expert opinions (if necessary).

Three types of scoring exist:

1. **Self scoring:** Each expert provides a self assessment in the form of a confidence level for each probability or answer provided for a subject.
2. **Collective scoring:** Each expert provides assessments of other experts, in the form of confidence levels.
3. **Entropy and information measures:** Scores for each expert are determined according to some rule of information reliability. These rules are discussed in the next chapter

H4 Models & mathematical backgrounds

§4.1 Entropy, (relative) information and calibration

When applying scoring rules, there are two kinds of properties you take a look at: entropy and calibration.

Let $p = (p_1, \dots, p_n)$, be a probability distribution over the outcomes x_1, \dots, x_n . The **entropy** of p then is:

$$H(p) = -\sum_{i=1}^n p_i \ln p_i$$

The entropy is always non-negative; it is 0, when some p_i is equal to 1, and $\ln n$, if all p_i equal $1/n$. Entropy is a good measure of the degree to which the mass is spread out, and provides a measure for the lack of information in the distribution p . When other things are equal, one should prefer the advice of an expert with low entropy.

Information represents the degree to which an expert's distribution is concentrated, relative to some decision maker-selected background measure.

The **information** of p , is the negative of the entropy:

$$I(p) = -H(p)$$

Entropy is a dimensionless measure for lack of information in finite distributions. When trying to find the entropy of a continuous analog f of the joint distribution above, there is no wholly satisfactory generalization for continuous probability densities. $H(f)$ behaves differently from $H(p)$. There is no universally accepted measure for information for continuous distributions. Cooke^[2] discusses a few practical solutions, which I won't mention here. However, we can use relative information, which is discussed further on in this paragraph (in contrast with entropy, relative information does have a generalization for probability densities).

To give a clear interpretation to the term calibration, think of an expert giving the same probability mass function p as above, for a large number n of unrelated uncertain quantities. By observing the true values for these quantities we generate a sample distribution s with s_i equal to the number of times the value i is observed, divided by n . An expert is well-calibrated if the true values of the quantities can be regarded as independent samples of a random variable with distribution p . This means that the difference between s and p should be no more than what is expected in the case of independent multinomial variables with distribution p . Therefore, well-calibrated must be translated into the statistical hypothesis H : "the uncertain quantities are independent and identically distributed with distribution p ".

So, **calibration** measures the statistical likelihood that actual experimental results correspond with the experts' assessments. Less formally, calibration is the probability that the divergence between an expert's probabilities and the observed values of the seed variables might have been arisen by chance. A low calibration says an expert's assessment isn't statistically supported by the set of seed variables.

The difference between s and p can be measured by **relative information**:

$$I(s, p) = \sum_{i=1}^n s_i \ln \left(\frac{s_i}{p_i} \right)$$

The relative information is always non-negative, and $I(s, p)$ equals 0 when $s = p$. Relative information reaches its maximum value when the entropy equals 0.

Large values of $I(s, p)$ are critical to the statistical hypothesis previously defined. The degree to which data support this hypothesis H , can be interpreted as the probability under H of observing a difference in a sample distribution s' at least as large as $I(s, p)$:

$$P_H(I(s', p) \geq I(s, p))$$

This probability can be used in statistical tests (see Chapter 5 about models) to test if calibration is significant.

Useful when conducting such tests, is the following fact: if p is concentrated on a finite number of integers M (which contains all observed values). Then as n goes to infinity, the quantity $2nI(s, p)$ becomes chisquare distributed with $(n-1)$ degrees of freedom. This can be shown by expanding the logarithms in the formula of the relative information via a Taylor series, and retaining the dominant terms^[2].

So, summarized: a good expert will show high calibration and high information (or low entropy).

§4.2 Scoring rules

As emphasized earlier, there are different types of scoring rules. These can be used to weight experts. A good scoring rules gives weights which:

- reward low entropy and good calibration;
- are relevant; and
- are meaningful to each individual variable, and the combination.

These conditions can be granted by using relative information and calibration.

H5 Models to combine expert's opinions

§5.1 Bayesian models

Bayesian models require that the user supplies prior probability distributions and process expert assessments by updating these distributions via Bayes' theorem. Many Bayesian models have been proposed during the past decennia in literature, but few have been applied. An exception is the model of Apostolakis & Mosleh^[2].

Let e experts give estimates x_1, \dots, x_e for an unknown quantity X . The decision maker starts with a prior density over X , and updates this with the information provided by the experts.

Bayes' theorem now reads:

$$p(X|x_1, \dots, x_e) = \text{const} * p(x_1, \dots, x_e|x)p(x)$$

If the experts are independent, this reduces to just the multiplication of the conditional probabilities. So only these have to be determined. Apostolakis & Mosleh provide two models for doing this. One of these models is the additive model:

$$x_i = x + \varepsilon_i$$

An expert's assessment is a combination of the true value and an additive error term. These parameters have to be chosen by the decision maker.

Under the assumption that the decision maker's prior $p(x)$ is normally distributed with mean μ_{e+1} and standard deviation σ_{e+1} , the conditional probabilities are normal with:

$$E(X|x_1, \dots, x_e) = \sum_{i=1}^{e+1} w_i (x_i - \mu_i) \text{ and } Var(X|x_1, \dots, x_e) = 1 / \sum_{i=1}^{e+1} (\sigma_i^{-2})$$

Proofs for these are rather difficult, and can be found in literature^[2].

The weights are:

$$w_i = \sigma_i^{-2} / \sum_{i=1}^{e+1} (\sigma_i^{-2}) \text{ and } \mu_{e+1} = 0$$

§5.2 Physical scaling models

These models are designed for estimating relative intensities of psychological stimuli (for example: taste). The models take as input qualitative data in the form of paired comparisons. A paired comparison is a judgement of which of two events is more likely. As output they give quantitative estimates of probabilities. The translation from qualitative to quantitative requires some strong assumptions.

Denote the number of experts with e . These experts have to compare objects $A(1), \dots, A(t)$. It is assumed that true values of these objects exist and that each expert e has an internal value $V(i, e)$ for $A(i)$. An expert judges an object $A(i)$ more probable to $A(j)$ is expressed as: $V(i, e) > V(j, e)$ ($= *$).

One physical scaling model is the **Bradley-Terry model**.

This model assumes that the probability ($*$) for all experts equals:

$$r(i, j) = \frac{V(i)}{V(i) + V(j)} \text{ with } \sum V(i) = 1$$

It is also assumed that each expert determines his view independently of the others. A good way to solve this system is to use the empirically observed proportions of experts to obtain maximum likelihood estimates of the $V(i)$. These estimates cannot be written in closed form. But it can be shown that they rely on the equation^[2]:

$$V(i) = \frac{\alpha(i)}{\sum_{j, j \neq i} n \{V(i) + V(j)\}^{-1}},$$

where $\alpha(i)$ denotes the number of times that $A(i)$ is preferred by some expert to some other $A(j)$.

§5.3 Classical model

The principles discussed in paragraph 3.3 and chapter 4 have been applied in the **classical model**. This is a performance based weighted average model, for combining experts' judgements. The weights are derived from experts calibration and information performance. This model is opposed to the weighting model where all experts are equally weighted.

The performance on calibration and information is measured at **seed variables**.

Seed variables are variables whose true values are unknown to experts when giving their opinion, but whose values are known post hoc. The performance of the experts on the seed variables is taken as indicative for the performance on the variables of interest in the study. This is a fundamental assumption of the classical model: Often, seed variables are not important to the study itself, but only to measure calibration performance of experts.

Seed variables serve the next three objectives^[12]:

1. To measure experts' performance as subjective probability assessors.
2. To enable performance-optimized combinations of expert distributions.
3. To evaluate and validate the combination of expert opinions.

The classical model combines calibration and information to yield a combined result (**score**) with the following characteristics^[8]:

1. The weights in the classical model are proportional to the product of statistical likelihood and information.
2. Calibration is more important than information; information serves to make a distinction between experts who perform equally well on calibration.
3. The maximum expected score for an expert is reached on the long term, if an expert gives his true opinion.
4. An expert is related to a statistical hypothesis, and the seed variables measures to what degree that hypothesis is supported by experimental data. If the likelihood score is below a certain cut-off point, the expert gets weight zero. The value of the cut-off point is determined by optimizing the calibration and information performance of the score.

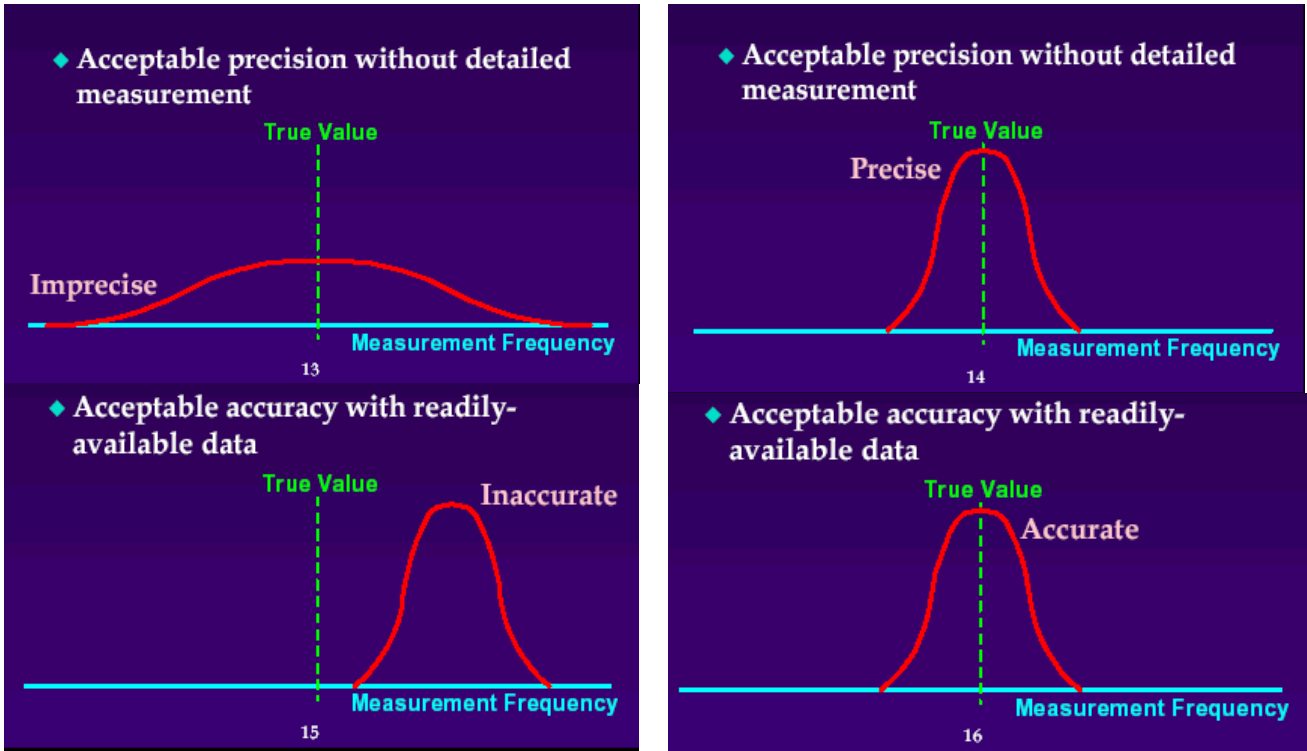
Weights for expert e are determined according to:

$$w(e) = C(e) * I(e) * 1_{\alpha}(C(e))$$

with $1_{\alpha}(C(e))$ the indicator function to test whether the calibration is significant.

Goossens et al.^[8] made some studies to compare the performance of the classical model with that of the equal-weighting-model. It is very hard to prefer one of the methods. Roughly it can be stated that they give similar calibration, but the classical model offers slightly better results for information performance. However, the results strongly depend on the number of experts used, the number of seed variables, and the robustness of the results against seed variables and experts.

The properties of the classical model meet the conditions, posed by Thompson^[11]: acceptable accuracy with readily-available data, or acceptable precision without detailed measurement. These conditions are illustrated by the figures below:



Figuur 4: Thompson's conditions

H6 Conclusions

In this paper I tried to give a comprehensive view of the history of the field of expert opinion.

Reasoning with uncertainty is difficult. Reasoning with uncertainty means dealing with subjective probabilities. Not everybody is an expert. Even though somebody can have substantive expert knowledge, he can perform poorly in assessing subjective probabilities, due to unfamiliarity with quantifying uncertainty. It is very important to train experts.

There are many mistakes about the use of expert opinion in practice. First of all: yes, there are many poorly conducted expert opinion studies. But these studies are mainly unstructured. Structured expert opinion studies can offer many possibilities and advantages.

During the past ten years the concept of structured expert opinion elicitation has been formalized, with better mathematical foundations.

The conclusions regarding the use of structured expert opinion are:

1. Experts' performance as assessors of subjective probabilities is not uniform: there are significant differences in their performance.
2. Performance based combination (the classical model) generally outperforms the equal weight combination, and offers better results than most unstructured or ad hoc methods.
3. A combined expert opinion may be satisfactory, even though the individual experts can perform poorly.

The conditions required for an expert opinion study depend greatly of the kind of study to be conducted: the size and the complexity of the study. The number of experts can vary between one to twenty; the amount of time between one day to one year. Factors determining the resources needed, are: travel time, training given to the experts, level of documentation, etcetera.

But my expectation is, that in the coming years, expert opinion will get a growing attention form all kind of social fields, because of the growing pressure on goverment and companies to know and manage their risks.

Bibliography

- [1] M.G. Stamatelatos - *RISK ASSESSMENT AND MANAGEMENT, TOOLS AND APPLICATIONS*, presentation for the NASA
- [2] R.M. Cooke - *EXPERTS IN UNCERTAINTY-OPINION AND SUBJECTIVE PROBABILITY IN SCIENCE* (1991)
- [3] H. Kahn - *THE YEAR 2000-A FRAMEWORK FOR SPECULATION* (1967)
- [4] Illinois Institute of Technology - *THE DELPHI METHOD*, <http://www.iit.edu/~it/delphi.html>
- [5] A.K. Rantilla & D.V. Budescu - *AGGREGATION OF EXPERT OPINIONS* (1999)
- [6] R.M. Hogarth - *ON COMBINING DIAGNOSTIC FORECASTS-THOUGHTS AND SOME EVIDENCE*, International Journal of Forecasting (1989)
- [7] B.M. Ayyub - *METHODS FOR EXPERT OPINION ELICITATION OF PROBABILITIES AND CONSEQUENCES FOR CORPS FACILITIES* (2000)
- [8] L.H.J. Goossens, R.M. Cooke & B.C.P. Kraan - *EVALUATION OF WEIGHTING SCHEMES FOR EXPERT JUDGEMENT STUDIES* (1998)
- [9] T.J. Gordon - *CROSS-IMPACT METHOD* (1994)
- [10] Illinois Institute of Technology - *CROSS IMPACT MATRIX METHOD OF FORECASTING*, <http://www.iit.edu/~it/cross.html>
- [11] P.D. Thompson – *NATIONAL HIGHWAY INSTITUTE BRIDGE MANAGEMENT COURSE* (2000)
- [12] R.M. Cooke & L.H.J. Goossens - *PROCEDURES GUIDE FOR STRUCTURED EXPERT JUDGEMENT IN ACCIDENT CONSEQUENCE MODELLING*