EUROPEAN COMMISSION

Community research

PAMINA
Performance Assessment Methodologies in Application to Guide the Development of the Safety Case

(Contract Number: FP6-036404)

REVIEW OF EXPERT JUDGEMENT METHODS FOR ASSIGNING PDFs
MILESTONE (N°: M2.2.A.3)

Author(s): **Ricardo Bolado and Anca Badea**
Joint Research Centre, European Commission

Michael Poole
Nuclear Decommissioning Authority

Date of issue of this report : **30/09/2009**

Start date of project : **01/10/2006** Duration : 36 Months

Project co-funded by the European Commission under the Euratom Research and Training Programme on Nuclear Energy within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the partners of the [PAMINA] project	
CO	Confidential, only for partners of the [PAMINA] project	

Foreword

The work presented in this report was developed within the Integrated Project PAMINA: **P**erformance **A**ssessment **M**ethodologies **IN A**pplication to Guide the Development of the Safety Case. This project is part of the Sixth Framework Programme of the European Commission. It brings together 25 organisations from ten European countries and one EC Joint Research Centre in order to improve and harmonise methodologies and tools for demonstrating the safety of deep geological disposal of long-lived radioactive waste for different waste types, repository designs and geological environments. The results will be of interest to national waste management organisations, regulators and lay stakeholders.

The work is organised in four Research and Technology Development Components (RTDCs) and one additional component dealing with knowledge management and dissemination of knowledge:

- In RTDC 1 the aim is to evaluate the state of the art of methodologies and approaches needed for assessing the safety of deep geological disposal, on the basis of comprehensive review of international practice. This work includes the identification of any deficiencies in methods and tools.

- In RTDC 2 the aim is to establish a framework and methodology for the treatment of uncertainty during PA and safety case development. Guidance on, and examples of, good practice will be provided on the communication and treatment of different types of uncertainty, spatial variability, the development of probabilistic safety assessment tools, and techniques for sensitivity and uncertainty analysis.

- In RTDC 3 the aim is to develop methodologies and tools for integrated PA for various geological disposal concepts. This work includes the development of PA scenarios, of the PA approach to gas migration processes, of the PA approach to radionuclide source term modelling, and of safety and performance indicators.

- In RTDC 4 the aim is to conduct several benchmark exercises on specific processes, in which quantitative comparisons are made between approaches that rely on simplifying assumptions and models, and those that rely on complex models that take into account a more complete process conceptualization in space and time.

The work presented in this report was performed in the scope of RTDC 2.

All PAMINA reports can be downloaded from http://www.ip-pamina.eu.

Proposal/Contract no.: **FP6-036404**

Project acronym: PAMINA

Project title: **PERFORMANCE ASSESSMENT METHODOLOGIES IN APPLICATION TO GUIDE THE DEVELOPMENT OF THE SAFETY CASE**

Instrument: Integrated Project

Thematic Priority: Management of Radioactive Waste and Radiation Protection and other activities in the field of Nuclear Technologies and Safety

Task 2.2.A Parameter Uncertainty
Topic 5: Expert judgement techniques for assigning PDFs

Review of Expert Judgement Methods for Assigning PDFs

Due date of deliverable: 31.03.08
Actual submission date: 30.09.09

Start date of project: 01.10.2006
Duration: 36 months

[*]Ricardo Bolado, [*]Anca Badea and [+]Michael Poole
[*]JRC, [+]NDA
Revision: 3
This version is an update to Revision 2, completed on 15.08.08.

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)
Dissemination level

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

1.- Introduction 3
2.- The need for Expert Judgement 6
2.1.- Types of uncertainties 6
2.2.- The use of Expert Judgement 7
2.3.- Advantages and drawbacks of using formal expert judgement protocols 8
2.4.- Remarks about the use and interpretation of expert judgement 10
3.- A theoretical framework for expert judgement 12
3.1.- Kolmogorov's axioms 12
3.2.- The Classical interpretation of probabilities 12
3.3.- Frequentist interpretation of probability 13
3.4.- Bayesian interpretation of probability 14
3.4.1.- The Bayesian update of information 17
4.- Biases and the assessment of experts 23
4.1.- Cognitive biases 24
4.1.1.- Biases related to the sources of information 24
4.1.2.- Biases related to the causal interpretation of the world 26
4.1.3.- Simplifying strategies (heuristics) 28
4.2.- Motivational biases 33
4.3- The assessment of experts 33
4.3.1.- Calibration curves 34
4.3.2.- Scoring rules 36
4.4.- The performance of experts 38
5.- Expert judgement elicitation techniques and protocols 40
5.1.- Techniques to assess probabilities and probability distribution functions 41
5.1.1.- Techniques to estimate probabilities of events 41
5.1.2.- Techniques to assess probability distributions for uncertain parameters 45
5.1.3.- Techniques to assess multivariate probability distributions 49
5.1.4.- Helping experts to provide their judgements 51
5.2.- Expert judgement protocols 52
5.2.1.- The Stanford Research Institute (SRI) protocol 52
5.2.2.- The SNL/NUREG-1150 protocol 56
5.2.3.- JRC's KEEJAM protocol 61
6.- Combination of expert judgement 63
6.1.- General characteristics of expert judgement combination 63
6.2.- Group combination 65
6.2.1.- Total interaction group 66
6.2.2.- The Delphi method 66
6.2.3.- The nominal group method 67
6.2.4.- The protocol used by Nirex and the NDA in the UK 67
6.3.- Mathematical aggregation 69
6.3.1.- The linear pool 70
6.3.2.- Bayesian combination of expert judgement 72
7.- Conclusions 73
References 74

1.- Introduction

This document deals with *Uncertainty* and the different ways to manage it. Uncertainty is an essential feature present in human life. We will never know with absolute certainty what is going to happen in the future, e.g. how good the weather will be tomorrow. Nevertheless, human beings have learnt to live under conditions of uncertainty. This is possible thanks to the development of strategies, technologies and heuristics (such as different types of insurance, daily weather forecasts or investment funds) that aim to compensate for the negative effects of uncertainty.

The scientific and technological fields are also strongly affected by uncertainty. In many cases the solution to a given technical problem is not readily available due to uncertainty, for example not knowing the theoretical model that could be applicable, or unavailability of data for whatever reason. In the case of the construction of a Nuclear Power Plant (NPP) or of a Radioactive Waste Repository, which are facilities required by society, but which also have the capability to threaten public health, usually several alternatives (different designs) are available, and we should be able to compare, in a fair way, their advantages and disadvantages. In these cases, expert judgement is a useful, and unavoidable, tool with which to solve the problem, filling the gap between the incomplete and limited information available, and the need to solve the problem.

Expert judgement may be necessary to understand the real magnitude of the problem to hand, to develop appropriate system models and alternative models, and to interpret the meaning of available data and their application. Judgements may also be used to define the characteristics of different alternatives in a decision problem. In general, whenever solving an important technical problem or making an important decision is required, the most capable people (persons with well-known knowledge and proved experience) should be made available to provide their judgements.

Given the fact that the use of expert judgement is unavoidable in the scientific and technical fields, the key question to answer is: should the opinions of experts be given in an implicit and informal way or rather in a explicit and formal way? Broadly speaking, we could say that informal judgements usually treat posed problems in a global way, not getting deep into the details. On the other hand, formal judgements get into the details of the issue under study, usually decomposing it in smaller and simpler, more tractable sub-issues. Moreover, when judgements are obtained in a more formal, structured and explicit way, they may be more easily documented, to facilitate further review. This may become very important when the issue under study is subject to review by a regulatory body, different types of associations, ecologists or any other potential stakeholder.

Knowledge Psychology is the branch of Psychology dedicated to the study of perception, cognitive processes and elaboration of judgements. It has identified the existence of different inferential mechanisms, typically used under conditions of uncertainty, which can be misused

introducing biases in judgements. As a consequence, judgements given by subjects can be inconsistent with the information they are based on, not providing an accurate picture of the real subject's uncertainty.

Judgements obtained following an informal, implicit process may be of use in daily life to solve non-important problems and make fast decisions. This is not acceptable when we are confronted with important technological problems affected by uncertainty and with a potential high impact on our society. Under these circumstances we must use formal expert judgement processes and require the opinions of qualified experts in that field. Structured and documented processes to obtain expert judgement in a formal way are called *Expert Judgement Protocols*, and consist of several phases designed to:

- train experts to provide formal opinions,
- identify and minimise the effect of biases,
- define the issue to be assessed with no ambiguity,
- make available to the experts all the relevant information about the issue under study,
- check the rationality and consistency of the opinions given, and
- make a final verification, repeating the whole process if needed.

The development of protocols and techniques to obtain formal expert opinions has been a consequence of the interest of many public and private organisations and individuals in incorporating, in the best possible way, uncertainties in their studies and decision-making processes.

Most of the techniques and the oldest protocols described in this document have been developed for generic problems affected by uncertainty, and are applicable in any area of technology, science, business, economy, etc. Nevertheless, the nuclear safety field has also made remarkable contributions to the development of structured expert judgement.

This report has been written within PAMINA's RTDC2 (treatment of uncertainty), and specifically under Work Package 2.2, task A, topic 5 (task 2.2.A, topic 5). This activity consists of characterising the uncertainty about the solubility limits for some key chemical elements (Radium, Tin, Selenium, Uranium and Plutonium) in the near field of a generic Spanish Radioactive High Level Waste repository in granite. This document is expected to help experts participating in this activity understanding what expert judgement is, and a significant part of its contents will be used during the training sessions.

The need to make use of formal expert judgement protocols to characterise key uncertainties in risk assessments for nuclear facilities will be justified in the second chapter. In Probabilistic Safety Assessments (PSAs) of NPPs and Performance Assessments (PA) for High Level Waste (HLW) and Spent Nuclear Fuel (SNF) repositories, it is quite frequent to deal with very low probability, high consequence events, and with datasets for crucial information that are of very small size. Expert judgement is vital in such situations.

Chapter 3 introduces a theoretical framework, the theory of Bayesian Probability, within which probabilistic safety studies of industrial facilities make sense and may be developed. Only within this framework can the use of expert judgement by justified. Chapter 4 presents the study of systematic errors (biases) committed by experts when giving their opinions. Biases are classified according to their origin as knowledge biases and motivational biases. Their origins and undesired effects are described. Finally, calibration curves and scoring rules, which are the main tools to characterise the quality of experts' judgements, are presented.

The first part of chapter 5 describes the most often used techniques to elicit the opinions of experts in terms of probabilities and probability density functions (or equivalently cumulative distribution functions) and the techniques to eliminate, or at least mitigate, the effects of biases. The main expert judgement protocols that have been developed are introduced in the second part of chapter 5. Because of its pioneering character, we introduce firstly the protocol developed by the Stanford Research Institute (SRI) of the Stanford University. Then we describe a protocol developed at Sandia National Laboratories (SNL) during the mid 1980's. The protocol developed at SNL was applied extensively in the NUREG-1150 study, a remarkable step forward in the area of PSAs for NPPs. Some other protocols, as for example the KEEJAM protocol, developed at the Joint Research Centre (JRC) of the European Commission, are more briefly described. Chapter 6 discusses to the problem of combining the opinions of several experts. Conclusions are presented in chapter 7.

2.- The need for Expert Judgement

During the last decades, risk analysis has arisen as one of the most powerful tools to study, in a structured way, the possible effects of complex industrial facilities on people and environment. The most widely accepted approach to perform this type of analysis is the one proposed by Kaplan and Garrick (1980). These authors consider a risk analysis of any system as the systematic answer to the following three questions:

1. What can go wrong?
2. What's the likelihood that things go wrong?
3. What happens if things go wrong?

Providing a formal answer to these three questions requires describing risk through the use of a set of triplets

$$R = \{< s_i, p_i(\phi_i), f_i(y_i) >\} \qquad i = 1,2,...,N., \tag{2.1}$$

where

- s_i represents scenario *i* in the set of n scenarios considered.
- $p_i(\phi_i)$ is the probability density function (pdf) that characterises our state of uncertainty about scenario *i*.
- $f_i(y_i)$ is the pdf that characterises our uncertainty about the potential consequences induced by our uncertainty about the system parameters under the conditions of scenario *i*.

According to this scheme, the first step in the process is to identify the potential scenarios that could affect the system performance. In the case of a PSA for a NPP, this means identifying the relevant Plant Damage States (PDS) and the accident sequences that may take the NPP to such damage states. In the case of a waste repository it means identifying the normal evolution scenario and the alternative scenarios. The second step consists of estimating the probability of each scenario-PDS (or, preferably the multivariate pdf that describes our state of knowledge about the occurrence of the different scenarios-PDSs). The last step consists of estimating the possible (adverse) consequences for each scenario. These consequences are uncertain because of the uncertainty that we have about the system model and the parameters required for it. In the next section we deal with the types of uncertainties that arise in a typical analysis of a complex industrial facility, e.g. a NPP under accident conditions or a radioactive waste repository.

2.1.- Types of uncertainties

Uncertainties in the development of risk analyses can be divided into two broad categories: Aleatory uncertainties and epistemic (or lack of knowledge) uncertainties. Aleatory uncertainties

are usually associated with (chemical or physical) parameters with some inherent variability. Aleatory uncertainties arise when an experiment is repeated several times under equivalent conditions and the results obtained differ from one another. An example of a parameter affected by this kind of uncertainty is the time to failure of a canister. We could fabricate a number of canisters following the same procedure and under similar production conditions and put them under similar physical and chemical conditions and measure the time that they take to fail. In this case, the variability in the results comes from the set of physical and chemical processes involved in the fabrication and in the operational (experimental) phases. Increasing the number of observations (canisters & experiments) does not decrease the aleatory uncertainty, but will allow us to know with more accuracy the probability density function (pdf) for the time to failure of the canisters, i.e. the type of pdf and the parameters that characterise it. So, for example, if that time follows a Weibull distribution, increasing the number of observations will allow us to know more accurately the minimum failure time, the standard deviation and the shape parameter.

Epistemic uncertainties are related to the existence of lack of knowledge about the problem. This type of uncertainty affects not only parameters, but also models and scenarios. A parameter will be affected by epistemic uncertainty when it is not random, but we cannot measure it, either because it is impossible or because it is extremely expensive to do it. This type of uncertainty is completely different from the aleatory uncertainty. Parameters affected by aleatory uncertainty are fully described by their associated pdfs. In the case of parameters affected by epistemic uncertainty, we try to characterise our lack of knowledge about the parameter, and we do it through pdfs. Those pdfs summarise our opinions about what values the parameter under study could more likely or less likely be close to. Many parameters (coefficients) of models used in the area of severe accidents and in the PA of HLW repositories are affected by lack of knowledge uncertainty; they are not random, but we are unable to know their values, so we have to use pdfs to characterise them.

Epistemic uncertainty affects models. Sometimes, there are several models to describe the behaviour of the system; some of them describe the behaviour of the system under some circumstances and others under other circumstances. It is not clear at all how to consider this in the analysis. Some authors consider appropriate to assign probabilities to the different competing models and to run one of them or another one according to those probabilities. Other authors consider the right solution to build up a meta-model that includes, as sub-models, the different models and run either one sub-model or another one depending on the values sampled and which models fit better with experimental results under those circumstances. Under any circumstance, only validated models, or at least non-invalidated models, should be used. Parametric studies with possible models could be used to obtain an estimate of the influence of model uncertainty.

2.2.- The use of Expert Judgement

The US Nuclear Regulatory Commission (USNRC or NRC) ordered in the early 1970s the first large scale NPP Probabilistic Safety Analysis (PSA) with formal treatment of uncertainties, the Reactor Safety Study (reported as document WASH-1400). This study was performed by Professor Rasmussen's group and completed in 1975 (Rasmussen et al., 1975). WASH-1400

identified, among others, transients, small LOCAs and human errors as important contributors to the risk for a NPP. These three items were the key features of the Three Mile Island accident, which happened a few years later, in 1979.

Though WASH-1400 was highly appreciated due to its pioneering character and its integral treatment of uncertainty, it was also subject to a lot of criticism. In fact, due to its complex structure, the methods developed and the importance of the conclusions, and in order to obtain a non-biased opinion about it, USNRC appointed a second committee, chaired by another academic, Professor Lewis (University of California at Santa Barbara) to review it. The results of this review (Lewis et al., 1975) confirmed some of the criticism already received, most remarkably about the treatment of common cause multiple failures and the way uncertainty had been represented and propagated, even the way it had been interpreted. As a consequence of this, when in the late 1970s, USNRC launched a programme to assess the risks associated with the geological disposal of Radioactive HLW, developed at SNL, uncertainty and sensitivity techniques were very much on the focus.

In the mid-1980's, USNRC started a new study to assess the risk associated with five commercial NPP in USA. The final report containing the results of this study is known as the NUREG-1150 report (USNRC, 1990). Among the aims of this study was the generation of quantitative estimates of the uncertainties affecting the risks computed, in order to avoid some of the criticism that were levelled at the WASH-1400 report. To achieve this objective, the uncertainty and sensitivity techniques developed at SNL were used. Additionally, expert judgement was also extensively used in order to obtain reasonable estimates and uncertainty measures for important parameters. Expert judgement techniques were used when limited information was available about the parameters. Since then, expert judgement has been considered an important technique in safety studies for NPPs and radioactive waste repositories.

2.3.- Advantages and drawbacks of using formal expert judgement protocols

Epistemic uncertainties must be certainly estimated in order to make a defensible risk assessment of any industrial hazardous facility. The only problem is to determine how such estimates should be obtained, either following structured, formal and well-defined processes or using more or less informal procedures. Bonano et al. (1990) stress the following advantages of using formal procedures:

1. Improved accuracy of expert judgements: This is because psychological biases are openly dealt with, problems are defined and communication is improved.
2. Well-thought-through design for elicitation: The procedures that are used in a formal expert judgement process are designed specifically for the problem being faced. The design relies on the knowledge concerning expert opinion, previous studies that have used expert judgement, and knowledge of the problem domain. Careful planning of the

process can substantially reduce the likelihood of critical mistakes that will render information suspect or biased.
3. Consistency of procedures: The participants follow the same procedures throughout a study and across related studies.
4. Scrutability: Documentation is a mandatory step of almost any formal procedure, which helps to ensure that various reviewers and users of the findings can understand and evaluate the methods and insights of the study.
5. Communication: Establishing a formal process helps to provide for reference documents useful in communication and external review. A formal process also encourages communication and understanding among experts and analysts about the problems studied and the values assessed.
6. Less delay: Projects have been delayed because critical judgements were not carefully obtained or documented, and a formal expert judgement process had to be designed and conducted before the project moved forward, DOE (1986), USNRC (1990).

Nevertheless, any structured process may also impose some restrictions, such as

1. Resources: There are costs in designing and implementing a formal process. Documentation is often more extensive with a formal process, and more resources are thus required.
2. Time: The time to establish and implement a formal process may be significantly greater than that required for an informal process. Scheduling of participants from external organisations adds a layer to the effort that is not present in an internal, informal process.
3. Reduced flexibility: Ongoing changes to the study are more difficult. If it is necessary to redo part of the study, re-enacting the expert judgement process may be cumbersome and expensive.

Taking into account these restrictions, it is not always justified to develop formal expert judgement procedures. Formal expert judgement procedures are expensive in terms of time and budget; they should only be implemented when the advantages exceed the drawbacks. In the NUREG-1150 study, many parameters whose uncertainties were originally intended to be estimated via formal expert judgement were eventually studied via informal expert judgement. Bonano et al. (1990) remark on the following reasons to justify the use of a formal expert judgement process:

1. Unobtainable data: Where extensive, non-controversial data directly relevant to a problem are lacking and unobtainable and existing data must be supplemented with judgements, it may be worthwhile to obtain judgements from experts using a formal elicitation process.
2. Importance of the issue: Formal methods are most appropriate when the expert judgements will have a major impact on the study and improvements in the quality of the judgements are then most worthwhile. Important issues also draw the most scrutiny. A formal methodology promotes documentation and communication and should be

employed when the issue studied is apt to receive extensive review and criticism or when the findings will be widely disseminated.

3. Complexity of the issue: When a problem is complex, or when several experts are employed either redundantly or as a team, formal methods are appropriate. These methods can provide the structure so that all participants understand the methods used and apply procedures consistently.
4. Level of documentation required: The critical reviews that the study will undergo, the variety and types of users, and the uses of the information may also suggest whether a formal process should be instituted. In some studies, the expert judgements may be important findings and, perhaps, used in subsequent studies, so formal methods are needed.
5. Extent of the use of expert opinion: When expert judgements are used extensively in a study, formalisation of the collection and processing of that information is apt to be done most accurately, consistently, and efficiently using formal methods. Costs that are fixed regardless of the size of the effort, such as creation of forms, training, etc., may be spread over many assessments.

These five reasons are met by a PSA of a NPP and by a PA of a radioactive waste repository. USNRC also prescribes the use of formal expert judgement processes when the following conditions are met (USNRC, 1994):

1. The issues to be assessed are very important for the final result or for the regulatory process.
2. When the issues to assess require a multidisciplinary approach.

In the case of the PA of a radioactive HLW repository, USNRC staff (USNRC, 1996) also provided their opinion, based on the accumulated experience, about the conditions under which formal procedures should be considered (when one or more of the following apply):

1. Empirical data are not reasonably obtainable, or the analyses are not practical to perform.
2. Uncertainties are large and significant to a demonstration of compliance.
3. More than one conceptual model may explain, and be consistent with, the available data.
4. Technical judgements are required to assess whether bounding assumptions or calculations are appropriately conservative.

2.4.- Remarks about the use and interpretation of expert judgement

The pervasive existence of epistemic uncertainties in the modelling of complex industrial facilities demands the extensive use of expert judgement. Nevertheless, the use of expert judgement must not be indiscriminate. The main reason for using expert judgement is *the presence of epistemic irreducible or almost irreducible uncertainties*. If this is not the case, the use of expert judgement can be questioned. The USNRC has very clearly expressed this idea, in a document (USNRC, 1991 - SECY-91-242) about addressing uncertainties in the

implementation of the Environmental Protection Agency (EPA) HLW standards, in the following terms:

'The formal use of expert judgement in performance assessments is a complement, rather than a substitute, for other sources of scientific and technical information, such as data collection and experimentation' (page E-11).

So, neither available data or data reasonably obtainable, nor technical and scientific rigorous analysis should ever be substituted for expert judgement. The only way to make sure the rational and adequate use of expert judgement, in a process where its wide use is foreseen, is the implementation of a quality system that plans, implements, reviews and documents its use (Kotra et al. (1996)).

A major problem encountered when working in the area of expert judgement is the widespread overconfidence that affects most of the experts (see chapter 4). In most of the studies there is a significant number of issues that are strongly affected by uncertainty and may have a strong impact on the results. Sometimes, the widespread diffusion of an average value, with no associated uncertainty ranges, may produce the illusion of accuracy, which may threaten the search for more and better quality information. Under these circumstances, it is important to keep in mind the fact that the task assigned to expert judgement is to characterise the uncertainty, not to reduce it. Again USNRC makes the following very pertinent remark (USNRC (1991) - SECY-91-242):

'Expert judgement should not be considered equivalent to technical calculations based on universally accepted scientific laws or to the availability of extensive data on precisely the quantities of interest. Expert judgements are perhaps more useful when they are made explicit for problems in which site data are lacking, since they express both what the experts know and do not know' (page E-11).

It is also important to realise that the state of knowledge of an expert about an issue is always referred to a given time. As the expert collects new data, becomes aware of new theories, etc., his/her state of knowledge may change, either increasing or reducing his/her uncertainty about that issue. Moreover, as each expert has a given background, access to different information and different ways to interpret it, it is possible that their individual states of knowledge about a given issue may be very different. Discrepancies between experts are logical consequences of the uncertainty that affects issues that are studied by means of expert judgement.

3.- A theoretical framework for expert judgement

During the last decades several new theories have arisen to deal with uncertainties, such as Zadeh's fuzzy set theory, Shafer's belief functions, Dempster's upper and lower probabilities, later on unified in the Dempster-Shafer theory of evidence. Nevertheless, no one of them has been enough developed as to be adopted by scientists and engineers in daily activities. The theory of probability remains as the main tool to deal with uncertainties; in fact probability is the metric used to measure uncertainty. It is a well-established theory, widely known by scientists and engineers, who apply it systematically in many commercial and scientific projects.

Nevertheless, there are still some problems related to the correct interpretation of the meaning of a probability; what a probability actually means and what things can be attributed a probability. This is very important in order to determine how both types of uncertainties (aleatory and epistemic) may be characterised.

3.1.- Kolmogorov's axioms

The theory of probability may be considered as the theory of non-negative additive functions defined on sets, whose axioms, developed within the theory of measure, is due to Kolmogorov (1956). Let U be the sample space associated to a given random experiment, i.e. the set of all possible results of the experiment. Every subset $A \subset U$ is called an event. A probability P is defined as a real function that assigns probability $P(A)$ to event A and satisfies the following properties (Kolmogorov's axioms);

1. For each event A, $0 \le P(A) \le 1$.
2. $P(U) = 1$.
3. If $\{A_i\}_{i \in I}$ is a finite or infinite countable set of mutually exclusive sets, then $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, where I is the set of indices that goes through the whole set.

It is important to remark that this set of axioms provides the rules about the way to combine the probabilities of simple events to compute the probabilities of more complex events. Nothing is said about the way the probabilities of simple events are computed. In the following pages we will describe the three main attempts to connect probabilities and their axioms with the real world or, in other words, the three main interpretations of probability: Classical probabilities, Frequentist probabilities and Bayesian probabilities.

3.2.- The Classical interpretation of probabilities

The classical interpretation of probabilities was developed by De Moivre and Laplace (Laplace, 1951). Both authors consider that, given a random experiment whose possible results are n mutually exclusive and equally likely events, and if n_A of them present attribute A, then the probability of such event is

$$P(A) = \frac{n_A}{n} \tag{3.1}$$

Let us consider the probability of getting the result '2' as a result of throwing a die. If the die does not show any defect, because of symmetry reasons we may deduce that there are six possible equally likely results, what means that the probability of such event is 1/6 ($n=6$, $n_A=1$).

Probabilities interpreted in this manner fulfil Kolmogorov's axioms, though they do also show important drawbacks: they are completely useless when the number of possible results of the random experiment is infinite and when the concept of equal likelihood (biased die for example) is not applicable. This is the case when dealing with lack of knowledge uncertainties, what makes this interpretation of probability of no use in a risk analysis.

3.3.- Frequentist interpretation of probability

Von Mises (1957) is among the developers and most active promoters of this interpretation of probability. Under this interpretation, probability is the limit of a series of relative frequencies. Given a random experiment, repeatable many times under similar conditions, and if event A is one of the possible results, its probability is defined as

$$P(A) = \frac{n_A}{n} \tag{3.2}$$

where n is the number of times that the experiment is repeated and n_A is the number of times that event A occurs. It is assumed that this limit does actually exist. Probabilities interpreted in this manner fulfil Kolmogorov's axioms.

Under this interpretation, probabilities are only meaningful when a random experiment, repeatable under similar conditions, may be set. Consequently these probabilities may not be assessed for one-of-a-kind events. *This type of probability may be of interest to characterise aleatory uncertainties, but not epistemic uncertainties*.

In addition to not being applicable to epistemic uncertainties, two more shortcomings are usually mentioned about frequentist probabilities:

1. We cannot experiment infinitely; information about the relative frequency will always be limited.
2. The system under study could vary over time, so that relative frequencies may also change over time.

3.4.- Bayesian interpretation of probability

Bayesian methods have their origin in the work published posthumously by the mathematician and Presbyterian minister Thomas Bayes in 1763 (Bayes, 1958). In modern times, the development of Bayesian probabilities as a measure of uncertainty are due to Ramsey (1926), Savage (1954, 1962), Lindley (1965) and de Finetti (1964, 1974). The Bayesian interpretation of probability is based on three fundamental ideas: degrees of belief, coherence and exchangeability.

- **Degrees of belief**

The Bayesian interpretation of probability extends significantly the field of application of the theory of probabilities through the introduction of propositions. A proposition is an affirmation about the occurrence of a given event. 'All bodies are attracted by the earth' is a proposition. In the Bayesian framework, propositions and events are treated homogeneously; we may say that an event either occurs or does not occur and we may also say that a proposition is either true or false. We may say that two are mutually exclusive; equivalently we may say that two propositions may not be simultaneously true. *The probability of an event or of a proposition is a measure of the degree of belief about the occurrence of the event or about the proposition's truthfulness*. If A is a proposition, and H is the knowledge of a person, $P(A|H)$ represents the probability assigned by that person to proposition A, it is the degree of belief of that person about the A being true, conditional on all the knowledge of that person. If the person beliefs that A is true, then $P(A|H)=1$, if he/she thinks that it is false, then $P(A|H)=0$. Other values in the interval (0,1) represent intermediate degrees of belief about the veracity or falsehood of A.

Within the Bayesian interpretation of probability, the proposition or event whose probability is assessed is as important as all the information used to base that probability on. If two persons assign different probabilities to the same event, it is because of the pieces of evidence/ information used to base their assessments on are different, so that the probabilities assessed are $P(A|H)$ and $P(A|H')$, which can be different. According to Lindley (1965), if both persons share their knowledge via honest discussion and exchange of ideas and information, they would arrive at the same probability for the proposition or event: $P(A|H,H')$. Not all authors agree about this statement. Savage (1954) thinks that, even after sharing all the information available, two persons could disagree about the probability of the event. This is a matter still open to discussion and research.

The most intuitive and straightforward interpretation of the probability concept is, in the opinion of the authors of this document, the one in terms of bets originally introduced by Ramsey (1926). This authors is of the opinion that if a person attributes value p to $P(A|H)$, it means that, if that person were invited to participate in a bet such that he/she would get a reward S in case A were true, and no reward if A were false, pS would be the maximum quantity the person would be willing to bet.

- **Coherence**

From a Bayesian point of view, there is no 'true probability' for a proposition; each person assigns probabilities to events based on their knowledge and opinions. This subjective aspect of Bayesian probability brought de Finetti to affirm, in the preface of his well-known book 'Theory of Probability' (de Finetti (1974)), that *probability does not exist*. Nevertheless, the individual freedom to estimate probabilities based on our own knowledge and experience is not a licence to make arbitrary estimations. The theory of probability requires *coherence* in the assessment of probabilities, which means that assessed probabilities must fulfil Kolmogorov's axioms and must also have the transitive property. Coherence is the objective normative requirement that must be fulfilled by any assessor.

A person is not coherent when his/her preferences do not fulfil the transitive property (de Finetti (1974)). Let us assume the existence of three possible alternatives *a, b* and *c*, among which we may choose one. The relation $a<b$ means that alternative *b* is strictly preferred instead of alternative *a*. A coherent person would establish a given order of preferences, such as $a<b<c$. Had he/she chosen the order $a<b$, $b<c$, $c<a$, that person would be behaving as a non-coherent person. Assuming that preference system, the relation $a<b$ means that, had he/she been obliged to take option *a,* and being then given the option to drop *a* and taking *b*, that person would be willing to pay up to a given quantity, say *x*, to change to option *b.* The relation $b<c$ means that the person would be willing to pay up to a new quantity, say *y*, to switch from alternative *b* to *c*. The relation $c<a$ means that the person would be again willing to pay up to a third quantity, say *z*, to switch from alternative *c* to *a*. The way this person set his/her preferences, he/she would be willing to pay up to a quantity $x+y+z$ to go from alternative *a* to alternative *a*, which does not make sense. If we want to avoid situations that take us to lose with complete certainty something that we appreciate, we have to make sure the transitive property in our preference system. Lack of coherence in a set of probabilities, interpreted in terms of bets, takes us to set probabilities that take us with complete certainty to lose. In Bayesian scientific literature this situation is called 'Dutch book'.

Coherence is not granted at all when subjects assess probabilities. De Groot (1988) provides the following example. Consider a box containing 90 balls, 30 of them are known to be red and the rest are blue and green, but the number of each colour is unknown. One ball is to be chosen at random from the box and you win a prize if you guess the colour of the chosen ball. What is your choice, red or green? Most people prefer to choose red, because they know they will get the prize with probability 1/3. Now suppose you are allowed to guess two colours and you win the prize if the colour of the chosen ball is one of those colours. In this situation most people prefer to choose blue and green, again, probably, because they know they will win the prize with probability 2/3. But both decisions violate the principle of transitivity. Let us call *R, B* and *G* respectively to the events *'getting a red ball', 'getting a blue ball'* and *'getting a green ball'.* In the first situation, choosing red means that our assessed probabilities fulfil the following relations: $P(R)>P(B)$ and $P(R)>P(G)$. Depending on the preferences between blue and green balls, the global system of preferences could be either $P(R)>P(B)>P(G)$ or $P(R)>P(G)>P(B)$. Since $P(R)$, $P(B)$ and $P(G)$ do not change from the first situation to the second, choosing green

and blue instead of red and blue means that $P(G)>P(R)$, which is in contradiction with the system of preferences set under the first situation (transitivity violation).

This example shows clearly that not being careful when assessing probabilities may produce a set of non-coherent preferences. In next chapter we will discuss several mechanisms, some of them very subtle, which dramatically affect the procedures used by subjects to process information in order to make decisions, and that may cause them to get them wrong. In next chapter we will also describe techniques available to help subjects assessing subjective probabilities.

- **Exchangeability**

De Finetti (1974) introduced the concept of exchangeability in 1931. The reason to introduce the concept of exchangeable events, and later on the concept of exchangeable quantities, was the conviction that it is impossible to learn from independent events, observations or quantities. Let $x_1, x_2,..., x_n$ be a set of random quantities (variables) that follow a given joint probability distribution $p(x_1,x_2,..., x_n)$, that represent the degrees of belief of a given person about them. The marginal distribution of a subset of m elements of this set is

$$p(x_1,...,x_m) = \int p(x_1,...,x_n)\, dx_{m+1} \cdot ... \cdot dx_n \tag{3.3}$$

The distribution of a subset of not yet known variables $x_{m+1},..., x_n$, conditional on the quantities already known $X_1 = x_1,..,X_m = x_m$ is

$$p(x_{m+1},...,x_n | x_1,...,x_m) = p(x_1,...,x_n) / p(x_1,...,x_m) \tag{3.4}$$

If the quantities x_i are mutually independent, then

$$p(x_1,...,x_n) = \prod_{i=1}^{n} p(x_i) \tag{3.5}$$

so that

$$p(x_{m+1},...,x_n | x_1,...,x_m) = p(x_{m+1},...,x_n)\,. \tag{3.6}$$

So, no learning may be expected when working with mutually independent random quantities (the conditional and the non-conditional distributions are the same). The idea in the mind of Bayesian statisticians is that, if we are to learn from experience, there should be something in the predictive distributions - $p(x_1,x_2,..., x_n)$, that would allow us to obtain more information about future events as we obtain more and more data, in other words, there should be some implicit dependency among the studied quantities included in the probability law $p(x_1,x_2,..., x_n)$.

Since the concept of independence is too strong, de Finetti tried to find a new concept that would relax the conditions of independence enough as to allow as learning from experience. This concept is *exchangeability*. A set of random quantities with probability law $p(x_1,x_2,..., x_n)$ is said to be finitely exchangeable if

$$p(x_1,...,x_n) = p(x_{\pi(1)},...,x_{\pi(n)}) \tag{3.7}$$

where $\pi(1),..., \pi(n)$ is a random permutation of the first n natural numbers. An infinite sequence of random quantities is infinitely exchangeable if any of its finite subsequence is finitely exchangeable, or in other words, the joint distribution of any finite subset of that sequence does not depend on what quantities are included in the sequence but only on the number of quantities included. Exchangeable random quantities are those that occur in a random sequence not being relevant, from the point of view of the form of the joint probability function, the order they appear.

Consider the case of flipping a coin and the sequence of possible results obtained when this experiment is repeated many times. Let as call x_i the result of the *i-th* experiment, whose result may be either head (1) or tail (0). Let us call $p(x_i)$ to the distribution of x_i and $p(x_{j1},...x_{jn})$ and $p(x_{k1},...x_{kn})$ to the joint distributions of two sequences of size *n*; *{j1,...,jn}* and *{k1,...,kn}* are two subsets of *n* elements of the set of natural numbers. Then $p(x_{j1},...x_{jn}) = p(x_{k1},...x_{kn})$, which means that the results of coin tossing are exchangeable events or quantities. So, the probability of any sequence of n coin tosses is the same as the one of any other sequence where the number of heads is the same. So, $p(x_1=1,x_2=1,x_3=0) = p(x_3=1,x_5=0,x_9=1)$, which is definitely true, the probability of getting head in the first and second toss and tail in the third one is the same as the one of getting head in the first and ninth toss and tail in the fifth.

3.4.1.- The Bayesian update of information

From all what has been discussed so far, we may deduce that two interpretations of probability may be used in risk analysis and performance assessment: the subjectivist (Bayesian) interpretation for one-of-a-type events and propositions and the frequentist for repetitive events. Though effectively, both cases could be included in the first one (Bayesian interpretation) for, if information about relative frequencies is available, the requirement of coherence will force subjectivists to assign probabilities very close to the observed relative frequencies.

The Bayesian interpretation of probability makes Bayes' formula a powerful tool to update degrees of belief when new information is available about an event or a proposition. Let *H* be the knowledge of a person (expert), and let $\{z_i\}_{i\in I}$ be a partition of the sample space of events. The Bayesian probability attributed by an expert to a given event z_k is $P(z_k|H)$. The acquisition of a set of new evidence *H'* produces a change in the probability given by Bayes' formula

$$P(z_k|H,H') = \frac{P(H'|H,z_k)\cdot P(z_k|H)}{P(H'|H)}, \quad (3.8)$$

where $P(z_k|H,H')$ is the 'a posteriori' probability of z_k, $P(z_k|H)$ is the 'a priori' probability of z_k and $P(H'|H,z_k)$ is the likelihood of evidence conditional on the knowledge H and the occurrence of event z_k. $P(H'|H)$ is the probability of new evidence conditional on previous knowledge, which may be considered a normalising factor, since the sum of expressions like (3.8) over the whole partition must be 1 (equivalently, the sum of the a posteriori probabilities of all the partition elements must be 1). That probability is given by

$$P(H'|H) = \sum_i P(H'|H,z_i)\cdot P(z_i|H), \quad (3.9)$$

and may be ignored in any intermediate computation. So, equation (3.8) may be written as

$$P(z_k|H,H') \propto P(H'|H,z_k)\cdot P(z_k|H), \quad (3.10)$$

which means that the a posteriori probability is proportional to the a priori probability and to the likelihood of evidence.

Two remarkable results are obtained from (3.8 – 3.10). If the a priori probability of an event is zero, the a posteriori probability will remain zero, even though the evidence against it could be very strong. So, much care should be taken when providing a priori probabilities. Null a priori probabilities should be avoided, unless total evidence about the impossibility of the events or propositions under study is available. In English literature this is called Cromwell's statement[1]. The second result is related to the existence of strong evidence. In that case, likelihood will be completely dominant and the a priori probability will be almost irrelevant (a posteriori probability and likelihood will be almost equal). This is the case of large sample sizes, for which relative frequencies and Bayesian probabilities will be almost equal.

Suppose we suspect a coin is not balanced (probabilities of getting head and tail are different), then we toss it *n* times. Before starting the experiments we have no reliable information about the probability *p* of getting a head, so we choose a non-informative *prior* distribution $\pi_0(p)$, for example a uniform distribution between 0 and 1. The prior distribution describes our state of knowledge about the probability that we want to study. The chosen distribution means that we know nothing at all about *p*, that is why we consider any possible value as likely as any other one (uniform distribution) and all the values that a probability may take (from 0 to 1). Suppose that the result of the experiment is *r* heads and *n-r* tails. This empirical evidence is used in Bayes formula to update $\pi_0(p)$ in order to obtain a new (posterior) distribution for *p*

[1] Gentlemen, I beseech ye, think ye, in the bowels of Christ, that ye may be wrong. Sir Oliver Cromwell addressing Parliament around 1651.

$$\pi_1(p) \propto p^r (1-p)^{n-r} \cdot \pi_0(p) \tag{3.11}$$

where the likelihood associated to the empirical evidence is obtained from the well-known formula for a Bernouilli process. When *n* is large (strong evidence), the likelihood is almost null everywhere except in a small interval around *p=r/n*, where it reaches its peak. As an example, if *n=30* and *r=5*, we obtain the following (posterior) distribution for *p*

$$\pi_1(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} = \frac{\Gamma(32)}{\Gamma(6)\cdot\Gamma(26)} p^5(1-p)^{25} \tag{3.12}$$

Figure 3.1 shows $\pi_0(p)$ and $\pi_1(p)$ for this example. When the prior distribution is non-informative, the posterior distribution is exclusively determined by the data (likelihood and posterior distributions are equal). The posterior distribution is significantly different from 0 around *p=5/30≈0.17*. So, a coherent person, whose knowledge includes the observation of frequencies obtained from many experiments, will assign subjective probabilities close to the observed frequencies.

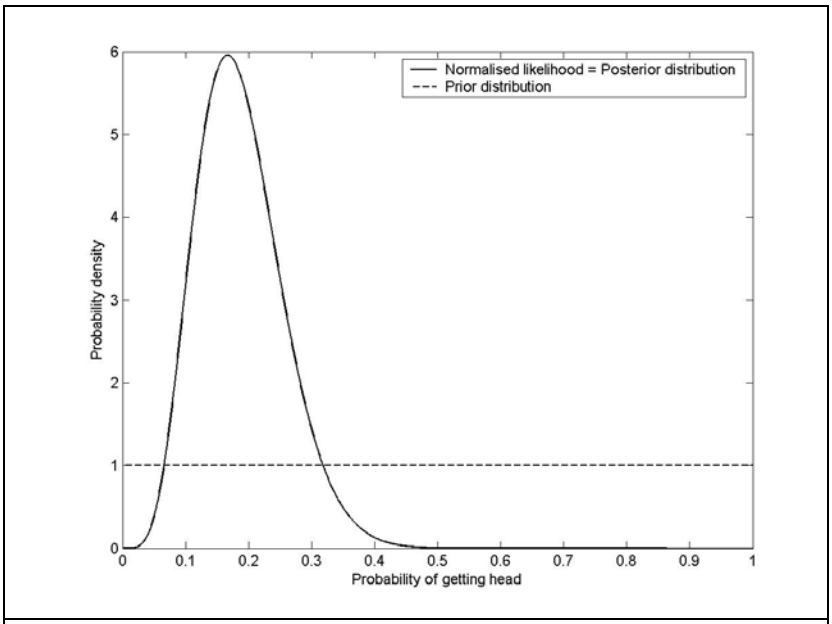

Figure 3.1.- A priori (prior) distribution, a posteriori (posterior) distribution for the unbalanced coin example. The likelihood distribution is the same as the posterior.

Bayesian inferential methods are most used under conditions of scarcity of data. The main steps of the formal process are similar to the steps of a classical inferential process: The selection of the probability model, the estimation of parameters and the diagnosis of the model. The main difference is in the estimation process, which is subject to the use of Bayes formula, as explained above. In the next paragraphs is an example of Bayesian estimation.

Let us assume a random variable X whose pdf is $f(X|\theta)$. This pdf is completely defined by the parameter θ, that is unknown and we want to estimate it. In order to start this estimation process, under the Bayesian framework, the parameter θ is considered as a random variable characterised through an a priori distribution $\pi(\theta|H)$. The a priori distribution provides information about the values the person/expert expects θ could likely take. In order to improve our knowledge about θ, we take a sample - evidence - $\mathbf{X}=(X_1,X_2,...,X_n)$, which will have $P(\mathbf{X}|\theta,H)=\prod_{i=1}^{n} f(X_i|\theta)$ as a likelihood function. Applying Bayes' formula provides the a posteriori distribution to be assigned to θ:

$$\pi(\theta|\mathbf{X},H) \propto P(\mathbf{X}|\theta,H)\cdot\pi(\theta|H), \tag{3.13}$$

which is a new pdf.

Let us assume the specific case of a Gaussian random variable X. Let us also assume that we do not know its mean, μ, though we know its variance, σ^2. Let us assume that, given our knowledge about it, we think that μ should have some value close to μ_0, let us also assume that μ could be, equally likely, larger or smaller than μ_0, and the further away from it the less likely. Under these conditions, a Gaussian a priori pdf for μ, with mean μ_0 and variance σ_0^2, could be justified. So that $\pi(\mu|H)\sim N(\mu_0,\sigma_0^2)$. Given a sample taken from the studied variable, its associated likelihood would be:

$$P(\mathbf{X}|\mu,H)=(2\pi\sigma^2)^{-n/2}\cdot e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}. \tag{3.14}$$

When putting this expression into (3.13) and after some computations, we obtain as an a posteriori distribution for μ

$$\pi(\mu|\mathbf{X},H)\sim N(\mu_n,\sigma_n^2), \tag{3.15}$$

where μ_n and σ_n^2 are

$$\mu_n=\frac{\frac{n\overline{X_n}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} \qquad \text{and} \qquad \sigma_n^{-2}=n\sigma^{-2}+\sigma_0^{-2}. \tag{3.16}$$

A priori, μ was considered to take values around μ_0, while after getting the information contained in the sample, values considered likely are those around μ_n. Additionally, the larger the sample size *n*, the closer μ_n and the sample mean, $\overline{X_n}$, will be (the larger n the larger the information contained in the sample is, while the a priori information remains constant). σ_n^{-2} is the accuracy of the estimation (the sum of the accuracy of the a priori distribution, $1/\sigma_0^2$, and the sample

accuracy, n/σ^2). The larger the a priori knowledge and the larger the sample size are, the larger the accuracy (the smaller the variance) of the a posteriori knowledge about μ. Figure 3.2 shows the normalised likelihood, and the a priori and the a posteriori pdfs assuming the following data: $\sigma^2 = 2$, $\mu_0 = 14$, $\sigma_0 = 2$ and $\mathbf{X} = (-3, 15, 23, 8, 13, 17)$. As previously described, the mean of the a posteriori distribution, $\mu_n = 12.43$, is between the mean of the a priori distribution, $\mu_0 = 14$, and the point where the likelihood function reaches its maximum, $\overline{X_n} = 12.17$. With the classical Maximum Likelihood Method, the estimate would be: $\hat{\mu} = \overline{X_n} = 12.17$.

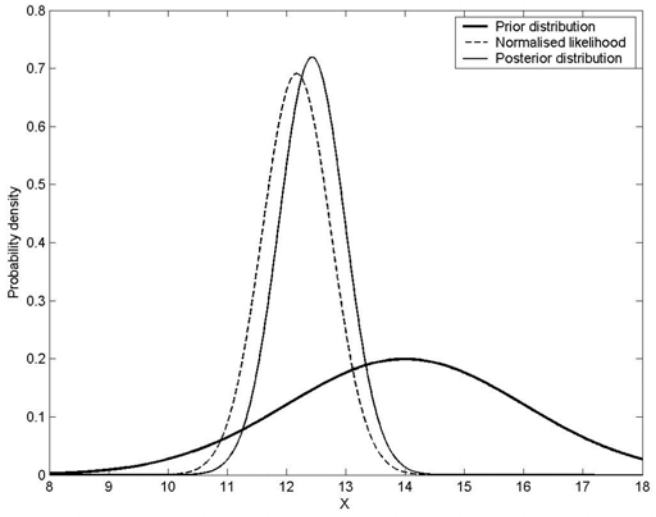

Figure 3.2.- A priori distribution, a posteriori distribution and likelihood function for the example of estimation of a Normal (Gaussian) random variable.

The validity of this estimation method is supported by:

1) its consistency with the way human beings learn from experience, and
2) by its convergence to the results provided by the Maximum Likelihood Method when the sample size increases (after analysing the first expression in (3.16), the reader may check that when sample size increases, the mean of the a posteriori distribution converges to the sample mean, which is the estimator of μ provided by the Maximum Likelihood Method), independent of the election of the a priori pdf, except in the aforementioned case of null a priori probabilities.

As a summary of this chapter, we should keep three ideas in mind about the Bayesian interpretation of probability and the framework that it provides to assess uncertainties. The first one is the extension of probabilities to propositions, the problem of measuring uncertainty about one-of-a-type events is now tractable, lack of knowledge uncertainties are measurable in terms of degrees of belief. The second one is that degrees of belief are personal, *the probability of an*

event does not exist any more, what does exist is the probability that a person attributes to an event, but degrees of belief are not a licence to arbitrariness, they must be based on a coherent preference system and must be based on empirical evidence if available and on high quality rationale. Finally, the third one is the unavoidable use of Bayes formula as the tool to incorporate new empirical evidence in our judgements and opinions.

4.- Biases and the assessment of experts

Human beings, in their daily activities have to make judgements about parameters and events. A judgement is an inferential cognitive process used to reach conclusions about the quality or quantity of unknown things, which is based on the available information. According to the existing vast literature (Rohrbaugh, 1979) about cognitive processes, the creation of judgements is developed in three steps:

1. Assignation of a given relative importance to each source of information.
2. Development of a specific functional relation between each piece of information and the final judgement.
3. Use of a specific method to integrate all the dimensions of the problem.

Moreover, experimental research in knowledge psychology has reached two solid conclusions: human beings have a limited capability to process information and show a strong tendency to perceive and interpret the surrounding world in a causal manner. The limitations of human beings to process information have their origin in four main deficiencies:

1. Our perception of information is not complete, but selective, human beings acquire only a fraction of all the information they have access to.
2. We cannot process information in parallel, but sequentially over time.
3. We do not have the capability to perform computations in an 'optimal' way. Instead of this, we use simplifying strategies, called heuristics, in order to process available information.
4. Our memory capabilities are limited.

Furthermore, our tendency to interpret the world in a causal manner makes our performance as 'intuitive statisticians' really poor; we have real problems in identifying situations of randomness, and accept them as such, and in being able to apply basic results of the theory of probability and statistics.

It has been widely tested that these limitations introduce biases in the three steps of the judgement creation process, which produces lack of accuracy and bias. This fact obliges us to identify the mechanisms used by subjects to make judgements in order to design techniques to improve their quality. Ideally, these techniques should be compatible with the peoples' natural capabilities.

The biases so far mentioned are called cognitive biases or knowledge biases. In addition to these biases, there are others, called motivational biases, which have to do with the potential interests and attitudes (economic, ideological, etc.) with respect to the results of the judgement. Motivational biases must also be taken into account and they can be a sufficient reason to disqualify some experts to participate in an expert judgement study.

Under the Bayesian (subjectivist) interpretation of probability, the only one compatible with expert judgement as a source of probabilistic statements, the probability of an event is the degree of belief of a person in the occurrence of such event. The probability is not, under this interpretation, something inherent to the event, and any estimate is valid, provided that it fulfils the condition of coherence and that it is based on the best available information. Nevertheless, when the opinions of the subjects, experts in this case, are going to be used to solve a problem of social relevance, as is the case of a PSA for a NPP or the PA of a radioactive HLW repository, any observer could ask a question about the credibility and reliability of the results obtained in a given expert judgement study. This fact forces us to design methods to assess the competence and accuracy of experts.

4.1.- Cognitive biases

In order to introduce cognitive biases, it is convenient to group them around three issues related to the aforementioned problems that people face when dealing with making judgements. The issues to be dealt with in the following sections are: the biases related to the use of different sources of information, the biases associated to the differences between causal and statistical reasoning and the simplifying strategies (heuristics) used to solve problems.

4.1.1.- Biases related to the sources of information

When people are making judgements, it is of paramount importance to analyse the information available and to attribute relative importance to each source. This process may be biased by a wrong perception of the robustness of the different pieces of information or data, which is usually due to wrong assessments of their abundance, consistency and reliability. Other facts related to the way information is presented could also have a non-desirable impact.

Wrong perception of the abundance, consistency and reliability of data

A basic rule in statistics is that the more data are available, the more reliable the results obtained from them are. As a consequence of this rule, people trust more judgements based on abundant sources of information. Nevertheless, this rule must be taken with caution, since it is true only when the sources of information used are independent; if they are redundant or correlated, their validity to base judgements on them is certainly lower. At the limit, if the same information is provided twice, it does not help gaining more confidence in the conclusions derived from it. Using a second copy of the same scientific paper does not provide any additional knowledge. In general, the situation is not so clear; usually people have several different sources of information with different degrees of overlap. So it is important to be aware of the possible correlation and overlap between the different sources of information available, especially if, as a consequence, other alternative sources of information are ignored.

Consistency has to do with the degree of concordance between the pieces of information coming from different sources. Consistency means that there is no disagreement between the different sources regarding a given concept or value, but it does not mean anything else. In some cases, several scientific or technical references could be very consistent regarding the acceptance of a given idea or value, producing a feel of robustness. This would be justified if the original sources were independent. If all references take the reader to a single original paper, based on rough estimates or not very exhaustive experimentation, or to a paper setting some conjectures instead of a well-proven theory, consistency is not enough to guarantee forecasting capabilities. Some literature is available in the area of cognitive psychology that support the idea that people frequently discard conflicting pieces of information instead of incorporating them to their judgements. Hogarth (1975) finds this strategy, the psychological reduction of information via discarding conflictive pieces of information, as a very useful way to reduce the anxiety produced by having to face uncertain events. We may conclude that consistency is a sensible strategy to base judgements on it, provided that it is thoroughly checked and conflicting or alternative pieces of information are taken into account.

The *reliability* of data is extremely important when they are going to be used to build a predictive model. Reliability is a measure of how representative data are. Unreliable data have no predictive capability. Using data whose origin is not well known (way they were collected, accuracy of the measurements, etc.) to build a predictive model is very risky. In some cases the data could be reliable, but the model based on them could be unreliable, in that case its predictive capabilities would also be negligible. Consider that we have a set of data that include a given variable of our interest, which we would like to be able to predict in the future, and a set of potential explicative variables. If we build a regression model based on the method of least squares to explain the behaviour of the variable of interest as a function of the explicative variables, and we find that the coefficient of determination of the regression (R^2) is very low and no regression coefficient appears as significant in the statistical tests, the best estimate for the variable of interest is its sample mean.

Different interpretation of information according to the way it is presented

In addition to the quantity and the quality of the data, the way they are presented is also important, since it may produce different opinions about their importance. The order the pieces of information are introduced to the people may affect their opinions. Sometimes a *primacy effect* may be observed. This happens when subjects pay more attention to the first pieces of information they have access to. In other cases people pay more attention to the last data obtained (*surprise effect*). The time frame and the rate subjects receive information is also important. People are very much influenced by their first hypotheses and the more time they take to get further information the more information they need to change their opinions. The clarity of the information accessed is also very important; badly-structured information could have a negative impact on the process of making judgements.

4.1.2.- Biases related to the causal interpretation of the world

Human beings have real difficulties living in uncertain environments, where conflicting information is abundant, due to the anxiety that this causes. One of the ways people use to ameliorate this problem is to establish causal relations, so that they can make predictions based on those relations. As a consequence of this, people improve their capacity to find potential causes of events, not dedicating much effort to understand their occurrence in terms of probability. It is important to remind that causal and statistical relations are very different. The former are unidirectional, if *A* produces *B*, *B* does not necessarily produce *A*, while the latter are bidirectional, if *A* is related to *B* in statistical terms, *B* is also related to *A*. This characteristic of human beings is crucial, for their ability to give opinions in terms of probabilities depends on their capability to understand and use in a skilful manner statistical and probabilistic concepts.

Wrong interpretation of causal relations

Einhorn and Hogarth (1978) show that in order to find out the possible causes of a given event, people analyse the following four factors:

1. The context.
2. Imperfect indicators of causal relation (time order of events: cause first, then effect; covariation; space-time proximity and cause-effect similarity).
3. Possible ways to combine context and indicators.
4. Alternative causes and their likelihoods.

When people try to imagine possible scenarios, large, detailed and coherent series of events are frequently considered more likely than the individual events themselves. In this case, causal coherence is considered as a proof in favour of the likelihood of the series of events. In some cases, the strength of the causal way of thinking may be so large that subjects attribute a higher probability to the cause followed by the effect than to the effect itself (remember that an effect may be produced by more than one cause), or to the occurrence of cause and effect simultaneously. In the latter, we are facing the risk of mixing up the probabilities of the intersection of cause and effect with the probability of the effect conditional on the cause (Hogarth, 1980).

Another widespread problem derived from the causal perception of the world is the *confusion of the inverse*. In probabilistic terms this means mixing up $P(x|y)$ with $P(y|x)$. In order to understand the meaning and the importance of this confusion, consider the following example taken from De Groot (1988). A person wishes to know if he/she has a given sickness. Then, he/she undergoes a medical test. The result is positive: according to it he/she has the sickness. Let us call *x* the event 'having the sickness', and let us call *y* the event 'the result of the test is positive'. $P(x|y)$ is the probability of having the sickness conditional on getting a positive result to the test, while $P(y|x)$ is the probability of getting a positive result in the test conditional on having the sickness. The latter is the one that appears in medical literature, since the test is applied to many patients that suffer the

sickness (the objective is to detect the sickness in sick people, not in healthy people), and the result registered is the fraction of positive results (which is the main reliability measure of the test). Fortunately, in most of the test $P(y|x)$ is close to 1. Many persons are not able to distinguish between $P(x|y)$ and $P(y|x)$, which are equal only if $P(x) = P(y)$. Take into account that this condition is almost impossible to fulfil: if the test is good, most of the population should be sick to get $P(x) = P(y)$. The confusion has to do with the temporal order in which events *x* and *y* are perceived. If the test is good, the occurrence of *x* almost surely implies the occurrence of *y*, but the first reliable information we get (not just a guess) is the result of the test, and the causal relation is interpreted reversely: the result of the test is positive so the sickness is there.

Lack of capability to use intuitively statistical concepts

Two important biases related to the difficulties to think in a probabilistic manner are the *lack of sensitivity to base ratios* and the *lack of capability to update information.* In general, when we try to solve a problem, we have two types of information: general information, which usually is acquired before the problem is set, and specific information, which usually is acquired ad hoc, to solve the problem. Quite frequently subjects ignore general information, which is also called base information or a priori information, and most of the focus is put on the specific information. It is worthwhile to remember that in the absence of specific information, judgements should be completely based on general information. The right way to use both pieces of information is to combine them using Bayes' formula in order to get the a posteriori information. Nevertheless, Bayes' formula is not included in the set of inferential intuitive mechanisms used by people.

In order to see the effect of base information and the use of Bayes' formula, let us take again the sickness example. Consider that the reliability of the test is 90%, or in other words, that $P(y|x) = 0.9$. Consider also that the probability of a false positive (positive result of the test for a healthy person) is $P(y|\bar{x}) = 0.1$. If, as it was said in the first part of the example, the result of the test was positive, should we think that the probability of having the sickness is $P(x|y) = 0.9$? If the person has no additional evidence about the existence of the sickness, as suffering some of the symptoms or sharing on a daily basis objects with people infected (for infectious illnesses), we can be more optimistic. Suppose that the person asks for some statistical results at a healthcare centre and he /she finds out that the estimated fraction of the population that is affected by that sickness is 10^{-4}. Then, if we apply Bayes' formula we find that

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{P(y|x)P(x) + P(y|\bar{x})P(\bar{x})} = \frac{0.9 \cdot 0.0001}{0.9 \cdot 0.0001 + 0.1 \cdot 0.9999} = 0.0009 .$$

Certainly this probability is nine times larger than for any person chosen at random, since the positive result of the test provides some additional information, but it is still considerably smaller than 0.9, which was intuitively and wrongly attributed to $P(x|y)$. This example is subject to criticism, since the base information that should actually be used is a matter of concern.

Another frequent source of error, even among educated people with some notions of probability, is to mix up the mean with the median. Different experiments (see Beach and Swenson (1966) and Spencer (1961)) in which sets of numbers were displayed to subjects, and they were then asked for the mean, the median and the mode, showed that subjects' estimates had a high degree of accuracy when the distributions were approximately symmetric. Peterson and Miller (1964) did similar experiments drawing data from a population that was highly skewed. In this case the estimates for the median and the mode were quite accurate while the estimates for the mean were biased towards the median. This means that, in some cases, when subjects are asked to provide an estimate of the mean, they could be providing something closer to the median than to the mean. A possible reason for this mistake is the conceptual difficulty of assessing a sample mean, which involves a sum and a division versus the assessment of a median that involves only ranking the samples and counting, or the assessment of a mode that involves only the assessment of a region with the largest concentration of values. Similar results could be obtained if subjects were given histograms of the sample (or even plots of the theoretical pdf), estimating a mode from them is straightforward (the peak) and estimating the median is also relatively easy (divide the histogram in two parts with equal area) while estimating the mean would need computing an integral or at least a weighted sum.

Subjects seem also to have poor skills to interpret the meaning of the variance and to assess the variance of samples. Experiments developed by Beach and Scopp (1968) show that subjects systematically overestimate the variance of populations with large deviations with respect to the mode such as bimodal or multimodal populations, which usually come from mixtures, while they usually underestimate the variance of populations with small deviations with respect to the mean and when dealing with normal populations. The underestimation of the variance is a well-known bias called *overconfidence* and it is also linked to a heuristic called *anchor and adjustment*, which will be explained in next section. Subjects have also shown poor performance estimating either very large or very small probabilities. Usually small probabilities are overestimated while large probabilities are underestimated. It is also worthwhile to mention the lack of capability of people to estimate the simultaneous occurrence of events and, in general, the joint distribution of several variables.

Subjects also show a clear tendency to adjust their subjective distributions to the normal scheme. Winkler (1967) thinks that, in the case of people with some knowledge in mathematics, this is due to the stress that is put on this type of distribution in standard statistical programmes. However, Hogarth (1975) thinks that this is related to the tendency of people to reduce uncertainty. Symmetry, which is one of the most remarkable properties of the normal distribution, is one of the most powerful mechanisms available to human beings to reduce uncertainty. People also find easier to think in terms of symmetry than in terms of lack of symmetry.

4.1.3.- Simplifying strategies (heuristics)

When people have to make decisions under uncertainty, they must necessarily make hypotheses about the probabilities of involved uncertain events. People trust some simplifying strategies (heuristics) that turn the difficult task of assessing probabilities into a series of simple tasks. In

general, heuristics are quite useful and provide clear benefits to people, but sometimes they may produce systematic severe errors, as shown by Tversky and Kahneman (1974). These authors identified three fundamental heuristics: *representativity, availability, anchor and adjustment.*

4.1.3.1.- Representativity

Representativity is a heuristic used when subjects have to assess the probability that an event *A* belongs to a class *B.* When this heuristic is used, subjects try to figure out how much representative is *A* of *B*, or how much does *A* reminds us *B*. So when *A* represents *B* very well, the probability attributed to the fact that *A* belongs to class *B* is very large, otherwise this probability will be deemed small. This heuristic may be very useful in common life but serious wrong estimates may be derived from its blind used. It is very much related to the *lack of sensitivity to the sample size*.

Lack of sensitivity to the sample size

Representativity is typically used when assessing the probability of getting a given result when a sample is drawn at random from a population. The likelihood attributed to a possible result usually depends on the degree of similarity with the corresponding parameter(s) of the population. Consider the following example. The height of the members of a reference population follows a normal distribution with mean 175 cm and standard deviation 5 cm. Then, consider the following two experiments: 1) take a person at random from a population (unknown) and his/her height is 180 cm; 2) take 25 persons at random from a population (unknown) and their average height is also 180 cm. If subjects are asked which of the two samples is more likely to proceed from the reference population, many will answer that both are equally likely (or unlikely). This is completely wrong. People ignore the importance of the sample size. The fraction of the reference population whose height is 5 or more cm above the mean is almost 1/6. The case of the 25 people sample is quite different. The mean height of a sample of size 25 coming from the reference population is a random variable whose distribution is also normal with mean 175 and standard deviation 1 cm. 180 is 5 standard deviations far from the mean, which makes it a very unlikely result.

Tversky and Kahneman (1974) provide another example that shows the lack of capability of people to take into account the sample size even when it is stressed in the formulation of the problem:

> "A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 % of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 %, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 % of the babies born were boys. Which hospital do you think recorded more days?"

The results showed that out of the 95 interviewed subjects, 21 opted for the large hospital, 21 for the smaller hospital and 53 thought both hospitals recorded about the same number. In probabilistic terms, the number of boys born in the large hospital a given day follows a binomial law with parameters n=45 (number of trials – number of children born per day) and P=0.5 (probability of success – probability of getting a boy in a birth). The number of boys born in the small hospital follow the same type of distribution with n=15 and P=0.5. The probability that the

large hospital gets more than 60% boys a given day is the probability of getting more than 27 boys, which is 0.068, while the probability that the small hospital gets more than 60% boys is the probability of getting more than 9 boys, which is 0.151. Figure 4.1 shows both probabilities. As a consequence, the small hospital will record more such days. In general, in small samples deviations are smaller than in large samples in absolute terms, but they are larger in relative terms.

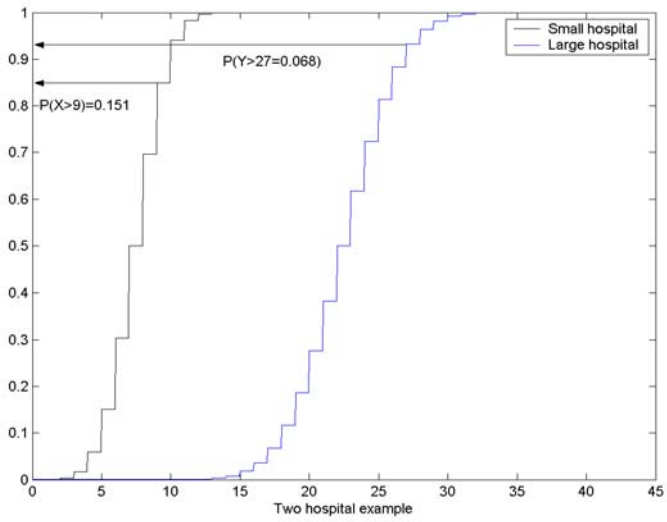

Figure 4.1.- Cumulative distribution functions for the number of boys born per day in the small hospital (black) and in the large hospital (blue).

Another example of lack of sensitivity to the sample size is what Tversky and Kahneman (1974) call the law of small numbers. This consists in expecting in small sequences of random events the same regularity as in large sequences. For example, people judge the string of coin tossing HTHTTH to be more likely than either the string HHHTTT or the string HTHTHT because they know that the process of coin tossing is random. The three sequences are equally likely to happen, however the first string looks more random that the other two outcomes in the opinion of many people.

4.1.3.2.-Availability

In some situations, the probability of an event is assessed according to how easy is to remember examples of that event. This is a very good strategy to assess probabilities since frequent events are more easily remembered than less frequent events. Nevertheless, the capability to remember the occurrence of events may be affected by factors other than the real frequency.

Wrong interpretation of the capability to remember

When availability is used to assess the size of a set, the elements that are more easily remembered look more abundant than those that are not so easily remembered. The capability to remember events or examples may be affected by facts such as recent personal experiences, the occurrence of catastrophes or the attention that communication media paid to the event. Lichtenstein et al. (1978) show an example in which several groups of well-educated American citizens were asked to estimate the number of people that died in America each year due to different causes. They were given a datum: each year roughly 50000 Americans die due to traffic accidents. Figure 4.2 shows the results of the experiment. The curve in that plot is the best fit for the geometric means. Had they been good estimators, the curve obtained would be the diagonal (number of estimated and actual number of deaths approximately equal).

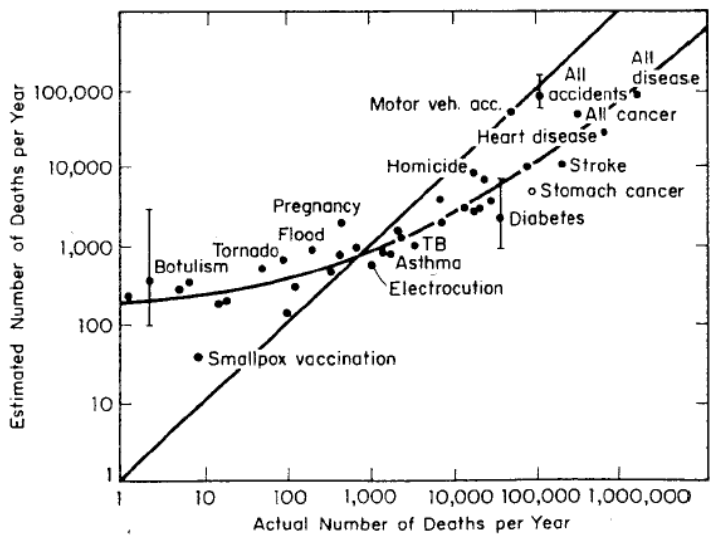

Figure 4.2.- Estimated number of deaths per year in USA due to different causes (geometric mean of individual experts' estimates) versus actual number of deaths per year (Lichtenstein et al.(1978)).

In the upper right we can see the results for cancers and coronary illnesses. These data are clearly underestimated. On the left we can see the results for rare illnesses and some natural catastrophes, whose numbers are overestimated. Only two persons died of botulism, but both cases appeared in all the newspapers of the country, which made them to be easy to remember. In the case of very frequent causes of death, people pay attention to them only when a relative or a friend is affected, which explains its underestimation. It is remarkable that for the less frequent causes of death the overestimation is of two orders of magnitude.

Lack of capability to imagine

The capability of subjects to imagine the conditions under which a given event may happen may affect the estimation of its probability. This is a frequent problem when assessing probabilities of events.

4.1.3.3.-Anchor and adjustment

When we try to assess the values that a parameter could take, usually we provide a first estimation (anchor), which is frequently interpreted as a measure of central tendency, such as the mean or the median. Later on we try to represent our uncertainty about it setting an upper and a lower bound. Both bounds are not usually set on themselves, but as distances to the mean (adjustment). This way of setting bounds for the uncertainty about a parameter is affected by the following problem, the first estimation has in general a very centralising effect, which quite often makes that the distances between the upper bound and it and between the lower bound and it be smaller than what would be expectable from the state of knowledge of the subjects. The centralising effect of the anchor has much to do with the lack of capability to imagine likely alternatives.

Tversky and Kahneman (1974) provide the results of an experiment that shows this bias. Two groups of people were asked to give their estimates about the percentage of African countries that belonged to UNO at the time the experiment was being performed. Each group was given by the authors a first estimate of that percentage: 10% to the first group and 65% to the second one. Though these values were selected on purpose, the authors said to the subjects that the value given to them had been chosen at random between 0 and 100. The median of the estimates given by the first group was 25% while the median provided by the second one was 45%, which shows clearly the effect of the anchor and the insufficient adjustment even when the anchor given to the subjects was supposed to involve no knowledge at all (random value). Lichtenstein's experiment (Lichtenstein et al. (1978)) also illustrates this bias. The authors repeated the experiment but in this case the datum was 1000 deaths due to electrocution. The effect of this new anchor was shifting the whole fitted curve downwards.

Overconfidence

When anchor and adjustment play a role in our judgements, the distributions provided are too narrow. This bias is called *overconfidence*. Overconfidence is easily detected when a calibration process is performed. These types of processes are discussed in section 4.3.

4.2.- Motivational biases

Motivational biases are related to the existence of a predefined attitude of the experts towards the expert judgement exercise to be performed. The adopted attitude may be legitimate or illegitimate. The most important motivational biases are

- Manager bias: This occurs when the value that the parameter under study could take is considered as a target instead of as a quantity whose uncertainty has to be characterised. Somehow, experts affected by this bias try to 'optimise' its value instead of estimating it.
- Expert bias: This is a possible consequence of the expert's reaction to be selected as an expert. If he/she has been selected is because he/she is expected to have real knowledge about the problem under study, which could be interpreted as having very little uncertainty. This attitude may produce a serious problem of overconfidence.
- Conflict of interest's bias: In some cases the expert may receive a part or even all his/her incomes from the organisation that is interested in performing the expert judgement exercise. Under these circumstances he/she could feel him/herself under the obligation of supporting the official opinion. In other cases, depending on the result of the exercise, the expert could either obtain a research project or start a new line of investigation or, in general, see his/her professional life modified depending on the results of the expert judgement exercise, which could modify his/her honest decision.
- Conservatism bias: It may be twofold. In some cases the expert could know what is the impact of the distribution selected for a parameter on the model results, and he could adopt the strategy of assessing a distribution that produces a conservative impact on the results, instead of trying to make an assessment as accurate as possible. In the second case, being aware of his/her potential overconfidence, he/she could try to widen his/her uncertainty ranges further than what fits his/her actual beliefs. In general this bias is related to risk aversion personal attitudes.

4.3- The assessment of experts

After obtaining the subjective probability estimates from experts, it would be convenient to assess the reliability and quality of these estimates. In order to do this, it is necessary to design a method that enables us to measure such quality, which is not an easy task if we bear in mind the nature of subjective probabilities. According to Winkler and Murphy (1968), the quality of an expert's opinions may be split in two parts: his/her knowledge about the subject under study (substantive expertise) and his/her ability to set his/her opinions in probabilistic terms (normative expertise). A good meteorologist could be very skilful at forecasting next day's weather (good substantive expertise) while an analyst could be more skilful at providing normal opinions in terms of probabilities (good normative expertise).

It is widely accepted that both kinds of expertise, substantive expertise and normative expertise, are needed to obtain high quality judgements. Certainly, obtaining the collaboration of outstanding experts in the matter under study is a must; the opinion of non-experts in technical matters is completely useless. In fact, some experimental studies (Merkhofer (1987)) show that substantive expertise partially counteracts overconfidence bias. Normative expertise does also help counteracting overconfidence bias for being familiar with making statements in terms of probabilities and considerably facilitates the process of making judgements. Winkler (1967) shows in an exhaustive experimental study that lack of normative expertise may introduce many biases in the experts' judgements.

Though the concepts of substantive expertise and normative expertise are very useful to introduce the problem of the quality of the assessments, they are not operational concepts. The most straightforward way to quantify the quality of the forecasts is to compare them with the empirical evidence available, which is not always possible. According to this criterion, good experts should attribute high probabilities to events that do actually happen while they should attribute low probabilities to events that do not happen. Two tools are available to measure the quality of experts: calibration curves and the scoring rules. In both cases the information used is either information about past events or about events whose occurrence may be checked in the near future.

4.3.1.- Calibration curves

An expert is deemed to be well calibrated when the event probabilities assessed fit the observed frequencies. So, a set of events whose probability estimate is 0.8 should happen roughly 80% of the times. This is the rule to build calibration curves. In most cases the questions used are almanac questions such as: how many inhabitants has a given country? What is the height of a given mountain? Or when did a historical event happen?

In order to build calibration curves for discrete events, events are grouped according to the similarity in the assigned probabilities. Suppose that n events have been assigned a probability p, and nt happened. If the expert is well calibrated then $p=nt/n$, while both values are different ($p\neq nt/n$) if the expert is not well calibrated. If this is done for each range of p, a plot like figure 4.3 may be obtained. A well-calibrated or ideal expert will produce a calibration curve that follows the diagonal quite closely; an underconfident expert, who shows too much uncertainty in his/her assessments, will provide a very large fraction of his/her assessment close to 0.5. The most frequent case is the overconfident expert, who provides quite more assessments too close to either 0 or to 1 than what would be desirable.

A similar calibration curve may also be drawn to assess the calibration of experts when they are assessing pdfs for continuous parameters. In that case, experts are supposed to assign pdfs to parameters whose real value may be easily obtained. With the information collected a plot like the one shown in figure 4.4 is drawn. In this plot we put estimated quantiles in the x-axis; in the

y-axis the fraction of assessments that the asked parameter was smaller than each given estimated quantile is drawn.

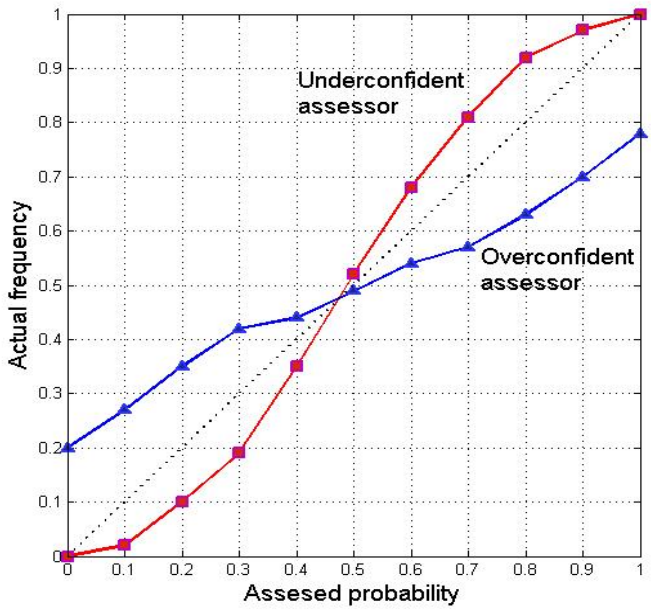

Figure 4.3.- Calibration curves for an underconfident and an overconfident expert based on estimates of probabilities of events.

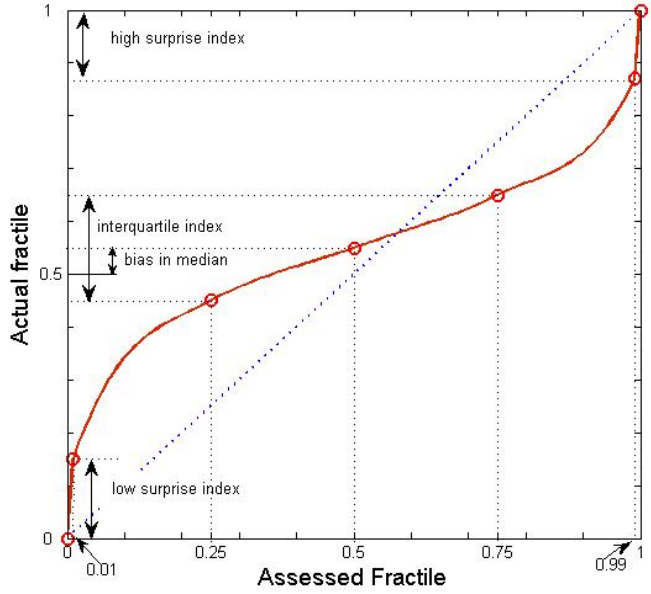

Figure 4.4.- Calibration curve for an overconfident expert based on estimates of pdfs for continuous random variables.

Figure 4.4 shows that 45% of the times the actual value of the parameter was smaller than the percentile 25% of the distribution estimated. Based on the information included in the calibration curves, three quantities of interest may be defined: the bias in the median, the interquartile index and the surprise index. The latter may be divided in the upper and the lower surprise indexes. The bias in the median is the difference between the fraction of times that the parameter took values below the assessed median and 0.5. The surprise index is the fraction of times that the assessed parameter was either smaller than the quantile 0.01 or higher than the quantile 0.99. As for the discrete event calibration curve, a well-calibrated expert should produce a curve close to the diagonal, with a bias in the median almost null, interquartile index close to 0.5 and surprise index close to 0.02. An overconfident expert, such as the one whose assessments have been used to draw figure 4.4, produces interquartile indices considerably below 0.5 (roughly 0.2 in figure 4.4) and surprise indices clearly above 0.02 (approximately 0.3 in figure 4.4).

4.3.2.- Scoring rules

Scoring rules were originally designed to encourage experts to make their assessments correspond with their judgments.

Consider a partition $\{E_i\}_{i\in I}$ of the sample space. Consider that an expert sets the distribution $\mathbf{r}=(r_1,r_2,...,r_i,...)$ for the partition when his/her actual judgement is $\mathbf{p}=(p_1,p_2,...,p_i,...)$. The expert says his/her true judgement only if $\mathbf{r}=\mathbf{p}$. A scoring rule is a function of the event that actually occurs and of $\mathbf{r}$: the expert gets a score (reward) $S_k(\mathbf{r})$ when event *k* occurs. Then, the expected score obtained by the expert is $E(S(\mathbf{r},\mathbf{p}))=\sum_{k\in I} p_k S_k(\mathbf{r})$. A scoring rule is said to be strictly proper if $E(S(\mathbf{p},\mathbf{p}))>E(S(\mathbf{r},\mathbf{p}))$ for every $\mathbf{r}\neq\mathbf{p}$. This scheme makes sure that the experts maximize their score only when they say their true judgments, provided that other necessary conditions such as the coherence of their judgments, the correct understanding of the method applied to obtain their opinions, the scoring rule itself and the linearity of their preferences with regard to the expected score.

Several strictly proper scoring rules have been designed (see Matheson and Winkler (1975), De Groot (1988) and Morgan and Henrion (1990)). The following ones are among the most popular ones

- The quadratic scoring rule: $S_k(\vec{r})=2\cdot r_k-\sum_{i\in I} r_i^2$
- The logarithmic scoring rule: $S_k(\vec{r})=\log r_k$
- The spherical scoring rule: $S_k(\vec{r})=r_k/\sqrt{\sum_{i\in I} r_i^2}$

In the case of two mutually exclusive events, the Brier score is the most used ($S_1(r)=-(r-1)^2$, $S_2(r)=-r^2$), which has been extensively applied in the field of weather forecasting. This scoring rule may be decomposed into three components, which respectively measure expert's knowledge, calibration and resolution (capacity to discriminate between different levels of probability). All

these scores may be generalized for continuous parameters, for example if the expert estimates a pdf $r(x)$ for variable *x,* the logarithmic score would be $S(r(x)) = \log r(x)$.

Matheson and Winkler (1975) have proposed a scoring rule for continuous distributions derived as an extension of binary events. Suppose *R(x)* is the distribution function given by the expert while *F(x)* is his/her real distribution function. Suppose *u* is an arbitrary value of *X* chosen by the analyst, which divides the range of *X* in two segments I_1 and I_2, as shown in figure 4.5. The scoring rule proposed is a function of the interval where *X* does actually takes its value and of the assessed cumulative distribution function *R(u)*:

$$S_I(R(u)) = \{ S_1(R(u)), S_2(R(u)) \} \tag{4.1}$$

where S_1 is applied if $x \in I_1$ and S_2 is applied if $x \in I_2$, so that the expected score is

$$E\big(S(F(u), R(u))\big) = F(u)\, S_1(R(u)) + \big[1 - F(u)\big] S_2(R(u)) \tag{4.2}$$

If S_1 is strictly proper $E\big(S(F(u), F(u))\big) > E\big(S(F(u), R(u))\big)$ for any $F(u) \neq R(u)$. Since the value selected by the analyst is not known a priori, in order to maximize the expected score, the expert has to provide his/her real opinion for all values.

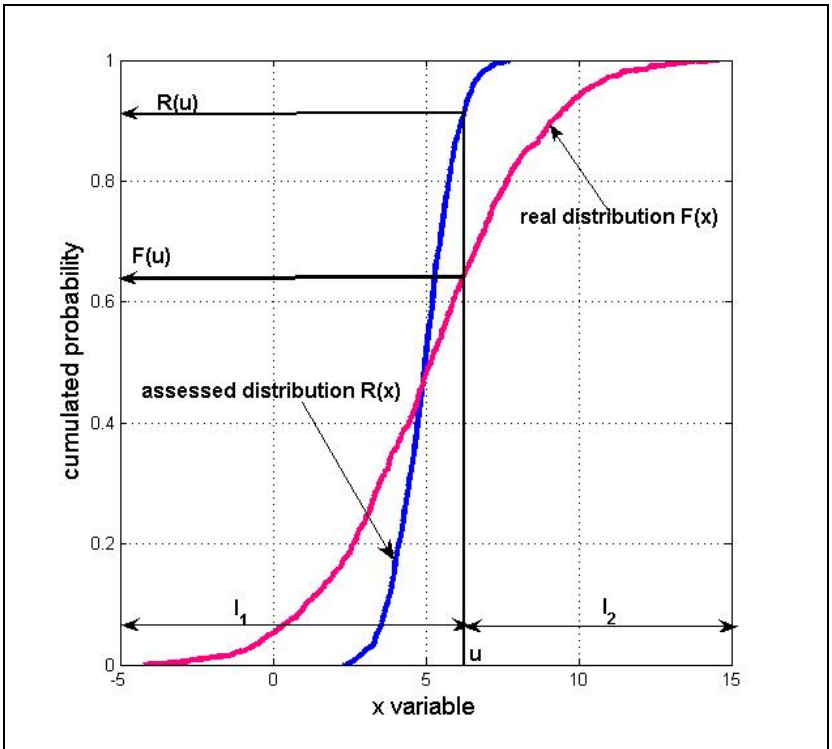

Figure 4.5.- Scheme used by Matheson and Winkler to develop their scoring rule.

This scheme may be independent of the selected value *u*. This may be achieved by integrating the scoring rule over *u* to get

$$S^*(R(\cdot)) = \int_{-\infty}^{x} S_2(R(u))\,du + \int_{x}^{+\infty} S_1(R(u))\,du \tag{4.3}$$

and the corresponding expected score becomes

$$E\left(S^*(R(\cdot))\right) = \int_{-\infty}^{+\infty} E\left(S(R(u))\right)du \tag{4.4}$$

From the very beginning of the use of scoring rules, they have also been utilized to assess the capability of experts. If an expert systematically gets better scores than others, he/she could be considered reasonably more reliable than the others. Nevertheless, this procedure to assess and motivate experts is also subject of criticism. According to Hogarth (1975), the use of scoring rules should be taken cautiously because of three reasons:

1. The expert is supposed to have a 'true' distribution in his mind, however they usually do not know exactly what their true distribution is.
2. The cost (psychological or otherwise) of making the assessment is not taken into account in the scoring rules, though certainly this is taken into account by subjects, who could consider the score not worthy of more effort.
3. Subjects lacking a sound mathematical background do not readily understand scoring rules.

Moreover, scoring rules seem not to be very sensitive to small deviations from the optimal strategy. Nevertheless, some authors such as Morgan and Henrion (1990) think that, even after considering the shortcommings of scoring rules, they remain as important tools to assess the performance of experts.

4.4.- The performance of experts

Most of the experimental results available in scientific literature about biases are based on studies done with students providing answers to general culture questions. This is why these results are considered with caution by some authors. In fact, some relevant authors like Lindley (1988) are not surprised about the doubts some psychologists have about people as probability assessors, mainly if their capability is estimated through questions like 'What is your probability that there are over 100,000 telephones in Ghana?'

Hogarth (1975) thinks that experts consider useful assessing probabilities only if two conditions are met: the issue to be solved via expert judgement should certainly be in the expert's field of expertise and the assessment to be provided by him/her should improve the state of knowledge about the issue more that any other reasonable alternative. This author warns about the validity of many experimental studies that have been done in the area of knowledge psychology, which in many cases posed trivial issues and subjects participating in the experiments had no specific expertise.

Mullin (1986) did a series of studies with experts in the areas of electromagnetic fields and hydrology, obtaining opinions about their respective areas of expertise and about general culture. The results of these studies showed a large difference between the assessments they did when they worked as experts and when they worked as non-experts. When they worked as experts they were quite more careful providing estimates, gathering information, identifying uncertainty sources and building models. Two of the main conclusions of these studies were that the results of experiments done with non-experts could not be directly extrapolated to experts, and experts were usually less overconfident than normal subjects when providing their opinions. Awareness of biases by formal training is probably the best that can be done to avoid them, instead of applying some not always well-justified de-biasing techniques.

5.- Expert judgement elicitation techniques and protocols

The need to quantify the uncertainties that appear in many technological problems, such as the Performance Assessment of a Radioactive Nuclear Waste Repository or the Probabilistic Safety Assessment of a NPP, demands the use of expert judgement. Not all uncertainties have an important impact on the final result of the study. Only when some non-reducible uncertainty has been identified as critical to the result of the study, expert judgement structured protocols should be used. Structured protocols are highly formal procedures designed to overcome the difficulties that arise in the process of obtaining the opinions of experts and enhancing the production of high quality expert judgements. Roberds (1992) has identified the following difficulties:

- Poor quantification of uncertainty: Experts can encounter serious difficulties in expressing uncertainty in a coherent manner. Usually this happens when experts are not well trained in probability and statistics, which is essential when we take into account the probabilistic nature of the questions they are asked.
- Poor problem definition: If the parameter whose uncertainty must be characterised has not been accurately defined, without any ambiguity, the problem posed to the experts will be biased from the very beginning.
- Unspecified hypotheses: Different experts may assume implicitly different underlying hypotheses. If these hypotheses are not clearly stated and explained to the analysts and to other experts, the conditional nature of the different assessments will not be realised. This and the previous difficulty would end up with the same problem: each expert could be solving a different problem.
- Uncorrected biases: Cognitive and motivational biases described in section 3 can seriously affect experts' judgements.
- Imprecision: The expert may be indifferent or insensitive over a range of values, which introduces fuzziness in his/her assessments. This may reduce the quality of the judgement.

These difficulties are the main reasons to refine the procedure used to obtain the opinions of the experts. The procedure should include steps and measures to eliminate or at least reduce the effects of the difficulties described above. This way, the judgements given by experts will more closely map their real knowledge and will take into account all the information available. Through the implementation of a formal expert judgement protocol analysts try to:

- Train experts in the coherent quantification of probabilities.
- Identify and minimise experts' biases.
- Define and document, with no ambiguity, the problem to be solved.
- Provide the expert with all the relevant available information.
- Obtain the opinion of each expert using the most suitable techniques, which may be different for different experts.

- Check and document the rationale and the coherence of each expert in his/her assessments.
- Make a final verification of the whole process, repeating it if deemed necessary.

Analysts play a fundamental role in formal expert judgement protocols for they have to understand what information experts are using to base their assessments on and how they are using it, so that they may identify the problems and biases that experts could face and they could take specific steps to minimise the effects.

In most of the protocols there is a phase in which analysts meet experts and obtain their opinions in terms of probabilities. In many cases, the outcome of that meeting will be the final solution given by the expert to the problem posed. Since experts will always be asked to give probabilistic statements, it is important to be able to provide them with a variety of means to map correctly their opinions. Many investigations have been developed in order to identify most efficient techniques to perform a correct mapping. Well-known analysts, based on all the experimental information and on their own experience, have selected a set of techniques as the best ones. In this chapter firstly the main elicitation techniques will be dealt with, then expert judgement protocols will be described in detail.

5.1.- Techniques to assess probabilities and probability distribution functions

In this subsection, a summary of the best available techniques to help experts providing their opinions is given. We will deal with techniques to assess probabilities of events and also to assess distributions for uncertain parameters, either univariate or multivariate. We will also discuss the advantages of training experts to counteract the effect of biases and to decompose problems in order to assist their analysis.

5.1.1.- Techniques to estimate probabilities of events

Subjects are not very used to making statements in terms of probabilities, with the exception of persons with some expertise in probability and statistics, and gamblers. Statements about probabilities are usually as imprecise as very likely, quite likely or not so likely. When an expert judgement exercise is planned it is because the events whose probabilities are under assessment are very important, some precision is expected from the experts, and preferably they are expected to be able to distinguish between low probabilities such as 10^{-4} and 10^{-5}. In some cases, even if they are able to distinguish between similar likelihoods, perhaps they are not able to do it in the typical scale between 0 and 1. That is why some techniques have been developed to make easier that translation from qualitative opinions to quantitative statements. The techniques may be classified as direct and indirect. Direct techniques may change the scale used to assess probabilities in order to adapt it to the capability of the expert. Indirect techniques use the

preferences of experts between different alternatives to derive probabilities and are very useful in helping subjects that are not familiar with the concept of probability.

5.1.1.1.- Direct techniques

The most straightforward technique of assessing a probability is to estimate it directly, but this is not always found as the most suitable way for some subjects; alternative other scales may be used. Two of these scales are the odds and the log-odds, which are

$$odd = \frac{P}{1-P} \tag{5.1}$$

$$\log - odd = Log\left(\frac{P}{1-P}\right) \tag{5.2}$$

The odd is the quotient between the probability of the events and the probability of the complement of that event. The log-odd is the decimal logarithm of the odd. Figures 5.1 and 5.2 show the scales used when those magnitudes are used to make probabilistic statements (the range of the odd scale is $[0,+\infty)$ while the range of the log-odd scale is $(-\infty,+\infty)$.

Some subjects are more capable of making probabilistic statements when they use the jargon of gamblers; they express their uncertainty about events via expressions such as ‘h against k in favour of the occurrence of the event’. When this jargon is used, the two following identities proceed

$$P = \frac{h}{h+k} \tag{5.3}$$

and

$$odd = \frac{h}{k} \tag{5.4}$$

where P is the probability of the event. When using this scale, a probability 10^{-3} would be formulated as ‘roughly 1 against 1000 in favour of the occurrence of the event’ (strictly speaking 1 against 999).

The use of odds and log-odds involves stretching the scale where probabilities are assessed, which should lead to an improved resolution. When the usual (linear) probability scale is used, it is not so easy to distinguish between probabilities such as 0.48 and 0.50, giving preference to ‘nicer’ numbers, 0.50 in this case. This enhanced resolution could help reducing both underconfidence and overconfidence, though experiments performed so far have not found significant differences in the quality of statements when using odds and probabilities. Bearing this in mind, the key reason to use any of these alternative scales is how comfortable experts feel when using them.

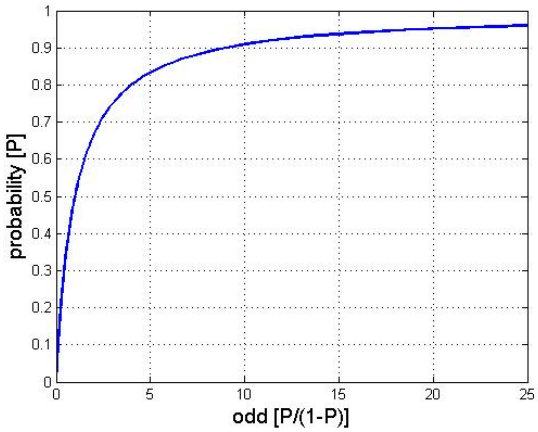

Figure 5.1.- Relation between probability and odd.

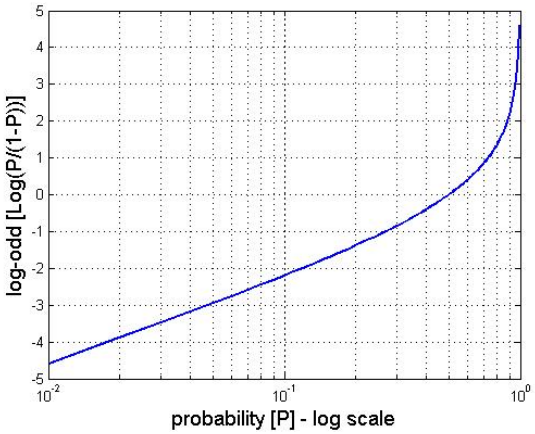

Figure 5.2.- Relation between Log-odd and probability (both in log scale).

5.1.1.2.- Indirect techniques

Savage (1954) believes that direct methods introduce large errors (discrepancies between assessed probabilities and their honest beliefs) because subjects are not good probability assessors; this is the reason why he proposes using methods that do not mention probabilities explicitly. Savage proposes a technique called 'certainty equivalence'. This technique consists in proposing to the expert a game: he/she will get an economic reward if the event he/she is asked the probability of occurs; otherwise he/she gets no reward. Then he/she is asked up to what quantity of money he/she would be willing to pay in order to participate in the game. The probability of the event is then assessed as the quotient between the maximum quantity of money he/she would be willing to pay to be allowed to participate in the game and the economic reward (this is the interpretation of probability suggested in chapter 3 as the most intuitive one). An equivalent alternative is to propose the expert to choose between the next two options: getting y euros if the event happens and getting nothing if it does not happen or getting a quantity x (not higher than y) independently of the occurrence or not of the event. The quantity x is increased until the point when the subject has no clear preference between both choices. The probability assessed would be the solution of the equation $x = P \cdot y + (1-P) \cdot 0$, which is $P = x / y$.

De Finetti (1974) does not trust this type of technique for assessing probabilities because, in his opinion, probabilities become biased by the attitude of the subject towards gambling. As an alternative he proposes a technique called the 'reference lottery'. The subject is confronted with two lotteries. In the first one he/she gets a reward x with probability P and a smaller reward y with probability 1-P, while in the second one he gets the reward x if the event A happens and y if it does not happen (see figure 5.3). The subject has to choose one of them. P is varied until when the subject has no clear preference between both choices, the value of P at that moment is the probability assessed by him/her for the event. The fact that the reward is the same in both lotteries is designed to remove biases produced by different attitudes towards gambling.

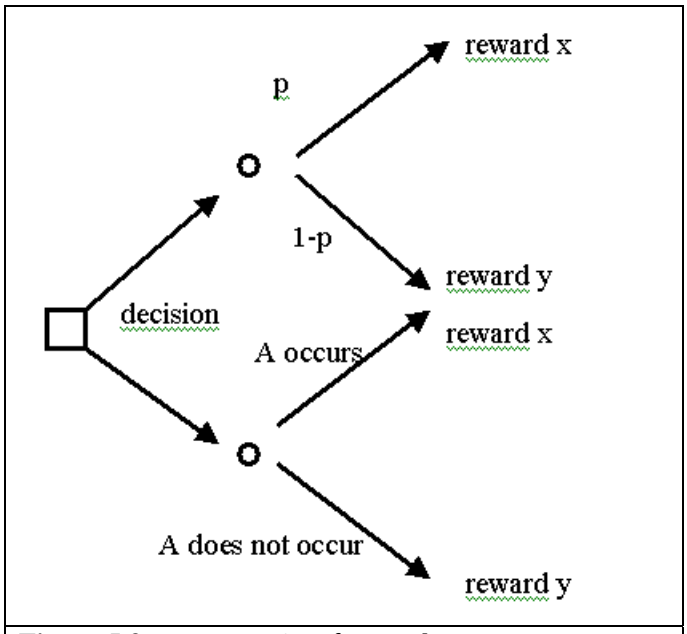

Figure 5.3.- De Finetti's reference lottery

Raiffa, due to the same reasons that brought De Finetti to propose the reference lottery, proposed the 'reference urn' method. The subject is encouraged to imagine an urn that contains balls with two different colours. The subject will be asked what ratio between the balls of both colours best corresponds with his opinion about the occurrence of a given event. Although this technique and the reference lottery seem appropriate to be used with subjects that do not have a good background in probability, many analysts find them tedious and difficult to apply in real expert judgement applications; they consider them useful only if either they are used to assess only a few probabilities or to introduce probabilistic concepts in the first steps of a protocol.

The last indirect technique that deserves mention is the probability wheel. This technique helps in visualising probabilities and is easier to use than the reference lottery and the reference urn. The probability wheel consists of two coloured paper circles (each with a different colour), a radial cut is done in each. Then we put one of the circles on top of the other in such a position that both radial cuts coincide. Then a part of the upper circle is put beneath the lower one. If we fix one of the circles and rotate the other one around the axis perpendicular to both through their centres, then the fraction that both circles overlap may be controlled by means of that rotation. Then a pointer that may freely rotate around the same axis is set. A scale may be painted on the edge of the lower circle so that the fraction P of area overlapped may be seen. In order to assess a probability, the subject is asked if he/she prefers to get a reward as a result of the pointer stopping on the overlapped area after being spun, or as a result of the occurrence of the event. The fraction of area overlapped is varied until the point that the subject is not able to make a choice between both options. The fraction of area overlapped at that moment is the estimate of the event's probability.

5.1.2.- Techniques to assess probability distributions for uncertain parameters

The techniques to help experts in assessing probability distributions may be divided into techniques targeting continuous distributions and techniques targeting discrete distributions. The techniques used to assess continuous distributions are based on the assessment of probabilities of discrete events such as 'the parameter value lies between this and that values', and the use of either interpolating or smoothing techniques.

5.1.2.1.- Discrete distributions

In the general case, experts need to assess the probabilities of n different possible values of the parameter under study. These n values may be considered as n different mutually exclusive events. When n is too large, say 10 or larger, it is convenient to group them in a smaller set. The first step is to ask the expert to rank them from the most to the least probable and to provide a rationale to justify such ranking. Later on, the individual probabilities are assessed. Usually the expert is asked firstly about the probability or the odd of the most likely value. In case the expert does not feel comfortable giving his/her assessment in terms of either probabilities or odds, other alternative techniques may be used.

The fact that the probability of all possible values has to add up to 1 makes it unavoidable to estimate the probabilities of at least n-1 values, the n^{th} may be deduced from the normalisation condition, yet it is advisable to understand the whole rationale of the expert, asking him the n probabilities, checking the normalisation condition later on. If the estimated probabilities do not add 1, it is always interesting to find out the reasons for such inconsistency. Normalisation may always be easily achieved, the normalisation constant k being obtained by solving the equation $k(\sum_{i=1}^{i=n} p_i) = 1$.

Lindley et al. (1979) have proposed another normalization technique based on Bayesian inference. Suppose that an expert assesses the incoherent distribution $\{q, \bar{q}\}$ for an event ($q + \bar{q} \neq 1$), though he/she has a true coherent distribution $\{\pi, \bar{\pi}\}$. If the analyst has an a priori distribution $P(\pi)$ for the possible values, then Bayes' formula may be used to incorporate evidence q in order to obtain the a posteriori distribution

$$P(\pi / q) \propto P(q / \pi) \cdot P(\pi) \tag{5.5}$$

The likelihood $P(q / \pi)$ represents the opinion of the analyst about the expert's normative knowledge, while the a priori distribution $P(\pi)$ may be computed as $P(\pi / A) \cdot p_A + P(\pi / \bar{A}) \cdot p_{\bar{A}}$, where $\{p_A, p_{\bar{A}}\}$ is the distribution attributed to events A and $\bar{A}$ by the analyst and $P(\pi / A)$ and $P(\pi / \bar{A})$ show his/her opinion about expert's knowledge about the issue under study. The true expert's opinion π, may be estimated as the expected value (mean) of the a posteriori distribution $P(\pi / q)$: $\hat{\pi} = E(\pi / q)$

This approach may be generalised to the case of a set of events/discrete values whose assessed distribution $\mathbf{q} = \{q_1,...,q_n\}$ is not coherent. Let us assume that each element q_i of $\mathbf{q}$ is normally distributed around the true corresponding expert opinion π_i of $\boldsymbol{\pi} = \{\pi_1,...,\pi_n\}$ with variance σ^2 and suppose that all those distributions are independent. Assuming that the analyst has a non-informative a priori distribution, then the a posteriori distribution is equal to the likelihood

$$P(\boldsymbol{\pi}/\mathbf{q}) = \frac{P(\mathbf{q}/\boldsymbol{\pi})\cdot P(\boldsymbol{\pi})}{\int_0^1 ... \int_0^1 P(\mathbf{q}/\boldsymbol{\pi})\cdot P(\boldsymbol{\pi})\cdot d\pi_1 ... d\pi_n} = k\cdot \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(q_i-\pi_i)^2\right\} \tag{5.6}$$

Then, the point estimates for each π_i subject to the restriction $\sum_{i=1}^{n}\hat{\pi}_i = 1$ will be

$$\hat{\pi}_i = q_i + n^{-1}\left(1-\sum_{i=1}^{n} q_i\right) \tag{5.7}$$

with the following variances

$$\text{var}(\hat{\pi}_i) = (1-n^{-1})\,\sigma^2 \tag{5.8}$$

where we can see that the improvement in accuracy due to imposing coherence is appreciable only when n is small.

5.1.2.2.- Continuous distributions

Two techniques are available to assess continuous distributions for uncertain continuous parameters: the quantile technique and the interval technique. The first of them is based on fixing probabilities and asking experts about the corresponding values of the parameter, while the second one consists in fixing values and asking about probabilities. Sometimes both techniques are combined in the same elicitation session. Other more complicated techniques, which require higher than average statistical skills can also be used, but some authors believe that these are not always well understood by subjects.

- **The quantile technique**

Given a random variable *X* with distribution function *F(x)*, its quantile *q* is the value x_q such that *X* takes values equal to or smaller than it with probability *q*, i.e.: $F(x_q)=q$. x_q may also be called percentile $100{\cdot}q$ %. The quantile 0 is the lower bound of the parameter under study and the quantile 1 is the upper bound. Quantile 0.5 is the median. The cumulative distribution function of *X* consists in plotting *q* versus x_q.

The application of these techniques begins by asking the expert about the upper and lower bounds of the parameter. If the expert is not able to provide such absolute bounds, then analysts ask about quantiles 0.99 and 0.01, or 0.95 and 0.05, depending on how comfortable the expert feels in estimating extreme values. It is important to combine these direct questions with others oriented to counteract overconfidence. These kinds of questions should encourage the expert to

think of conditions under which the parameter could take values outside the boundaries previously assessed and the likelihood of such conditions. It can be very beneficial for the analysts to help the expert in imagining such situations either by 'digging' into the problem or providing examples that came up in similar expert judgement applications. Einhorn and Hogarth (1978) note that experts frequently make good use of this type of help. In general, thinking of alternative conditions help experts in broadening their uncertainty ranges.

Once the upper and lower bounds, or some given extreme quantiles have been assessed, then the analyst asks about the median. The expert is asked about a value that could be equally likely smaller or larger than the actual value of the parameter. It is important to see how far this estimate is from the previously estimated bounds or extreme quantiles. If the estimate is more or less in the middle, this could mean that the expert is just taking the mean of the extreme values or considering some kind of implicit symmetry. Medians too close to some of the bounds could also indicate a poor original definition of the bounds. The expert should be asked about the reasons for providing that estimate for the median. Later on, quantiles 0.25 and 0.75 are addressed using similar questions. More quantiles may be assessed, but five quantiles are usually enough to draw an approximate cumulative distribution. The shape of the curve should be discussed with the expert in order to uncover potential inconsistencies. It is also useful to draw the pdf and show it to the expert, since features like the symmetry or lack of symmetry of the distribution are easier to see in the pdf than in the cumulative distribution.

It is important to address extreme quantiles and absolute bounds as the first steps of the elicitation session. In the first applications of this technique the bisection method became very popular. It consists of asking firstly for the median, asking then for the 25% and the 75% and so on. Nowadays this method has been abandoned because it was experimentally shown that asking firstly for the median converts this estimate into an anchor that quite frequently led to overconfidence.

- **The interval technique**

In order to apply this technique, the analyst selects some values and asks the expert about the probability that the parameter is located within the different intervals defined using those values. There are two types of intervals to define: open intervals and closed intervals. In the first case the analyst selects a point and asks the expert the probability of the value of the parameter being smaller (or larger) than it. In the second case the analyst selects two points and asks the expert the probability that the value of the parameter be inside the defined interval. If the expert finds difficulties in giving his/her opinions in terms of probabilities, the analysts can help him/her using indirect techniques.

In both cases (open and closed intervals), in order to avoid overconfidence due to introducing an anchor given by a first estimate about a central value, the analyst starts by posing questions related to very extreme values, that should correspond to quantiles such as 0.01,0.05, 0.95 and 0.99, or even to absolute bounds. Later on, some other interior points are selected, usually between three and seven, depending on the degree of accuracy required, and the analyst asks the

expert questions based on them. Each answer given by the expert must be supported by some rationale, and the expert should be confronted with conflicting data and with hypothetical situations not considered or mentioned in his/her rationale. When the closed interval option has been selected, another convenient task to do is to ask the expert to rank from the most likely to the least likely a number of intervals whose limits are set using the group of values selected. With the information obtained from the expert the cumulative distribution functions and the corresponding pdfs are built and results are discussed to check consistency.

- **Other techniques**

From a statistical point of view, the most straightforward method in characterising the uncertainty associated to a given parameter is to ask directly either the pdf or the cumulative distribution function. Two possible options are to provide him/her with graph paper or a specific software program to assist the expert in drawing the distribution. Both functions present advantages and drawbacks in characterising uncertainty. The cumulative distribution function is easy to assess due to its direct interpretation, allowing the fast identification of quantiles of interest. On the other side, the pdf is very useful for seeing characteristics related to the symmetry (or lack of symmetry) of the distribution, as for example the location of the mode. The use of specifically developed software to plot pdfs and cumulative distributions may be very useful to get very fast graphical feedback from qualitative statements related to the shape of the distribution.

Another method to generate the distributions, which should be used only with experts with a proved background in statistics, is to estimate directly the probability law (Gaussian, Weibull, Exponential, etc.) and its parameters (the mean and the standard deviation for the Gaussian distribution, the scale and shape parameters for the Weibull distribution, the expected value for the exponential, etc.).

Winkler (1967) has proposed two techniques, the equivalent prior sample (EPS) and the hypothetical future sample (HFS) to estimate the probability of an event using a Bayesian approach, see also Garthwaite et al. (2005). In both cases the beta distribution plays a central role in the estimation process. The author also reports about the difficulties encountered by experts in applying these techniques. Nevertheless, they could be of use if experts with enough knowledge in statistics participate in the exercise.

Hampton et al. (1973) reports about an indirect method developed by Smith (1967) to build pdfs called *psychometric classification.* This method consists in dividing the range of values that a parameter may take into several segments, say *n*, and ask the expert to rank them from the most to the least probable. Later on he/she is asked to rank from the largest to the smallest the differences in probability between the different segments. Pay attention to the fact that probabilities are not assessed at all; they and their n-1 differences are only ranked. Then, using a method suggested by Kendall (1962), firstly relative likelihoods between segments are assessed, then the actual probabilities are derived and the corresponding histograms are plotted. From the histograms, pdfs may also be derived. Experimental results confirm the precision and reliability

of this method, producing more spread distributions. Nevertheless, Morgan and Henrion (1990) think that the results obtained are more related to the data treatment process than to the assessments of the experts, in their opinion the method involves a recalibration of the experts. Hampton et al. (1973) consider that both rankings used, specially the second one, lack of any psychological and intuitive meaning, and are difficult to assess by experts.

- **The selection of the technique**

The results of experimental studies used to compare techniques are not conclusive, and in some cases they reach contradictory results. The final selection of the techniques should be based on the issue under study and the experience and preferences of experts. The most widely used techniques are the quantile and the interval technique due to the fact that subjects easily understand them. The other more mathematical techniques are not so frequently used.

There is some empirical evidence that the interval technique produces more spread and better distributions than the quantile technique. So an advisable strategy is to combine both in the elicitation sessions to check the consistency of the assessments generated. Assessments obtained using one of the techniques may be used to ask questions based on the other technique.

5.1.3.- Techniques to assess multivariate probability distributions

Given all the scientific literature available about the cognitive capability of human beings and their deficiencies, it is not a surprise to learn about the serious difficulties encountered in detecting dependency and correlation structures. Assessing multivariate distributions is a task that, in general, cannot be performed without the use of techniques to simplify it.

In the case of two discrete variables that may take the values $x_1,...x_n$ and $y_1,...y_m$, the easiest way to estimate their joint distribution is to estimate the marginal distribution of one of them, say $P(x_i)$, and the corresponding n associated conditional distributions $P(y_j|x_i)$ then, bearing in mind the relation between marginal conditional and joint distributions, we may get the joint distributions as $P(x_i, y_j)= P(x_i)\cdot P(y_j|x_i)$. Both, conditional and marginal distributions are assessed by means of the techniques described in section 5.1.2.1. In some cases, there could be some causal relation between X and Y, for example X could be the cause and Y the effect. Because subjects find it fits better their way of thinking, it is then convenient to assess $P(Y|X)$ (causal inference) instead of $P(X|Y)$ (diagnostic inference). Nevertheless, quite frequently subjects perceive first the effect, then the cause. In this case subjects could find easier to estimate $P(X|Y)$ and use Bayes' formula to estimate $P(Y|X)$. It is convenient under these circumstances to inform the expert about the possibility of being affected by the *confusion of the inverse* bias.

When the number of discrete variables is higher, the assessment of the joint distribution becomes more difficult, then it is convenient to use other techniques such as probabilistic influence diagrams, which provide an adequate framework to model problems where dependence and statistical relations between random variables appear simultaneously. A probabilistic influence

diagram (Barlow (1988)) is a directed acyclic graph that provides a graphical representation of the relations existing among the random variables that concur in a given problem (see figure 5.4). Variables are represented inside circles called nodes, which are connected with other nodes by arcs that denote possible statistical dependence. Associated with each node is a conditional probability function. Conditioning is only with the immediate predecessor nodes, which is indicated by the direction of the arrow. In figure 5.4 (a) Z is the predecessor variable of X and Y, which indicates that theses two variables are conditionally dependent on Z, and are conditionally independent with one another. Conditional independence means that $X|Z$ and $Y|Z$ are independent, which does not necessarily mean that X and Y be independent. The existence of no arc connecting two nodes that have a common predecessor means conditional independence. Nodes that are not targeted by any arrow are always unconditionally independent. In figure 5.4 (b) X and Y are unconditionally independent, while Z is conditionally dependent on both. Figures 5.4 (a) and 5.4 (b) lead to different ways to compute the joint distribution of X, Y and Z; in the first case $P(X,Y,Z) = P(Z)\cdot P(X|Z)\cdot P(Y|Z)$, while in the second case $P(X,Y,Z) = P(X)\cdot P(Y)\cdot P(Z|X,Y)$. The theory of probabilistic influence diagrams is highly developed and may be of help in analysing complex problems via expert judgement.

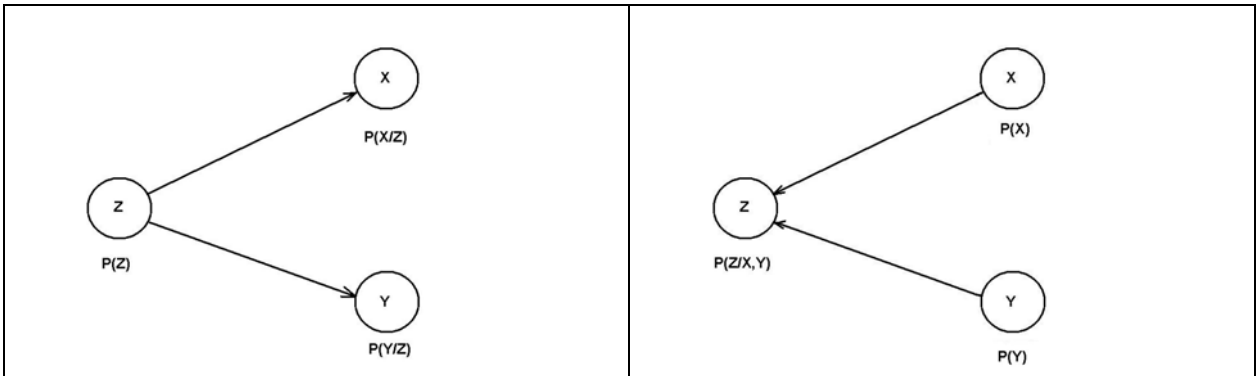

Figure 5.4 (a).- Influence diagram. X and Y are conditionally independent. Both are conditionally dependent on Z.

Figure 5.4 (b)- Influence diagram. X and Y are independent. Z is conditionally dependent on both X and Y.

When multivariate continuous distributions have to be assessed, the ranges of the variables need to be divided into several intervals and those intervals treated as if they represented individual values of discrete variables. Kadane et al. (1978) have developed a computer assisted method to assess multivariate normal distributions using the quantile technique. Chaloner and Duncan (1987) have also proposed a method to assess multinomial distributions. In some cases, knowing the linear dependence between two variables is the objective. This may be assessed through the estimation of their correlation. Gokhale and Press (1982) have proposed two methods based in assessing the probability of concordance and of exceedance to assess the prior distribution for the correlation coefficient of a bivariate normal variable.

5.1.4.- Helping experts to provide their judgements

Two activities have shown their importance and efficacy in helping experts providing their judgements: *decomposing the problem* and *training the experts in assessment techniques.*

5.1.4.1.-Decomposition

This strategy consists of decomposing the quantity of interest into others that are simpler to assess. Later on, after obtaining the distributions of the simpler variables, they are aggregated to get the distribution of the original quantity. A simple decomposition of a problem could be as follows: Suppose we want to estimate the number of cows in a given country. It could be interesting to decompose it as *number of cows* = (*number of inhabitants*) · (*annual per capita consumption of milk per year*) / (*average annual production of milk per cow*). This decomposition would be meaningful if obtaining the data corresponding to the disaggregated variables would be easier than obtaining the data corresponding to the variable of interest.

Judgements obtained via decomposition may represent more accurately the actual state of knowledge about the problem, because simpler assessments are frequently more accurate due to a better calibration. Though the usefulness of this strategy is supported by general principles in the area of cognitive psychology, a lot of experimentation is still needed to find out the set of the circumstances under which decomposition is beneficial and what is the optimal level of problem decomposition. Nevertheless, this strategy has become one of the most popular ones and is applied by analysts extensively.

Mosleh et al. (1988) reviewed a large quantity of expert judgement applications. They found that in many of these applications, decomposition was applied in a very coarse fashion, even in situations when it was completely meaningless. They point out three situations in which decomposition may be most effective:

- There is much uncertainty,
- A relevant theory exists for certain aspects of the problem under study,
- Different experts have information about different aspects of the problem.

Regarding the optimal level of decomposition, Bonano et al. (1990) show that it does not payoff to decompose the problem beyond a given optimal level. That optimal level is obtained as a balance between the number of assessments to perform, the complexity of the decomposition and the computational complexity of the aggregation. The analyst plays a key role in determining that optimal level. Mosleh et al. (1988) think that decompositions developed by the experts themselves, without the help of the analyst, are very interesting for they allow deeper insights about expert's rationale and help the experts to become more comfortable with the whole assessment process.

The computational aggregation of the assessed variables obtained via decomposition demands knowledge of the functional relation between them and the quantity of interest $y = f(\mathbf{x})$. If f is a

simple function the aggregation may be done analytically, otherwise Monte Carlo simulation is a simple way to do it.

5.1.4.2.- Training in assessment techniques

Overconfidence is probably the most serious threat to the quality of expert judgement applications and though different techniques perform differently, no one guaranties good performance. Morgan and Henrion (1990) reviewed a large number of experimental studies about the assessment of continuous distributions and the corresponding calibration curves. They found that the interquartile range varied between 20% and 40%, while the surprise index varied between 5% and 40%. As is known, they should approximately be 50% and 2% for a well calibrated expert. Most analysts believe that training experts in providing their judgements may help in reducing overconfidence and increasing accuracy. This happens because experts experience a feedback process that helps them to understand the different techniques and to improve the inferential process they use to generate their assessments. In fact, according to results of knowledge psychology, human beings are adaptive organisms that use different strategies to make their judgements according to the issue being assessed.

Obtaining feedback during training sessions and observing how the quality of judgements improves across a series of iterations is very interesting. Either scoring rules or calibration curves may be used. It is also convenient to have a short discussion with the expert after each assessment to discuss the problems encountered. Nevertheless, there is no empirical evidence about what should be the characteristics of a good training programme; we only have the experience of many analysts and the idea that it is necessary and useful. A problem related to training experts is that it is a time and budget consuming process.

5.2.- Expert judgement protocols

Several expert judgement protocols are available in the scientific literature. In this section we are going to describe the following ones: the Stanford Research Institute protocol (SRI protocol), the SNL/NUREG-1150 protocol and the Knowledge Engineering Methodology for Expert Judgment Acquisition and Modelling (KEEJAM protocol) developed by JRC-Ispra. Protocols may be classified according to what types of opinions are targeted in the elicitation sessions, individual opinions or group opinions. Protocols whose objective is obtaining group judgements are described in next chapter.

5.2.1.- The Stanford Research Institute (SRI) protocol

This is the first structured protocol developed to obtain individual expert judgements and can be considered as the precursor of most of the others. It was developed in the 1960's and 1970's by the Decision Analysis Group of the SRI (Stanford University). Other protocols, as for example Delphi, were developed earlier, but they are group opinion protocols. The protocol was originally divided in five phases, though it was further enlarged (Merkhofer, 1988), after the dissolution of

the Decision Analysis Group, with the inclusion of two more phases to deal with the combination of several experts' opinions and the discretisation of the assessed distribution. In this protocol, the process to obtain the opinion of the experts is considered as the joint task of an analyst and an expert. A description of the protocol follows:

- **Phase 1: Motivating**

The objective of this phase is a first contact with the expert in order to inform him/her about what is expected from him/her and to find out if there is any risk of encountering motivational biases.

Firstly, the analyst explains to the expert the general problem to be solved and the context in which his/her opinions are meaningful. The expert is informed about the importance that sensitivity analyses performed attribute to the parameter or event under study, he/she is also informed about the way the opinions provided will be used within the general problem. Then the expert is informed about the fact that the objective is not to forecast a single value but to characterise the uncertainty about the parameter of interest. This is very important, especially if during the conversation the analyst detects the possibility of facing either manager or expert bias.

The next step is to discuss openly with the expert about the possibility of being affected by some other motivational biases. If any of them is detected, the analyst has to take some steps to counteract them, as for example changing the rewards pattern for that expert or decomposing the quantity of interest. The strategy adopted may depend on the detected bias. In the case that motivational biases are so severe that they could damage the quality of the assessment, the expert could be discarded, though this is only done in extreme and rare cases.

- **Phase 2: Structuring**

The purpose of this phase is twofold. On the one hand, the objective is to decompose the quantity of interest as a function of several other variables in order to simplify the task of assessing distribution functions, on the other hand the objective is to obtain information about the way the expert approaches the problem, identifying implicit hypothesis that were unknown to the analyst and that could introduce bias into the assessment results. This phase may be divided into three steps:

1. To set an accurate definition for the parameter under study: The parameter under study must be precisely defined. An important tool to apply is the clairvoyant test. It consists of asking whether a clairvoyant would be able to provide the value of the parameter with no uncertainty. For example, asking the price of a UO_2 pellet in 2020 would not pass this test. It would be necessary to set explicitly the enrichment factor, the reactor type, the supplier, the currency and its reference year (euros of January 1st 2020), etc.
2. Study the possibility of decomposing the quantity of interest: Decomposing the quantity of interest may help in counteracting the effects of motivational biases. Working on low-level variables may help in disconnecting an expert's assessments from his/her personal

interests, which are based on high-level variables, or attitudes towards risk and uncertainty. The authors think that decomposing the problem helps experts in assessing joint probabilities, making assessing a conditional probability instead of the right joint probability less likely.

3. Explicitly set all hypotheses to be used by the expert: The objective is to uncover all implicit and explicit hypotheses considered by the expert in his/her assessment. A useful method of uncovering implicit hidden hypothesis is to ask him/her what he/she would like to insure against. In other words, if he/she were allowed to take insurance on certain events that could make his/her estimates wrong, what would be those events? Finally, the units to be used in the assessment for the quantity of interest or for any low-level variable that comes up in the decomposition process must be clearly stated. Experts should be allowed to choose the units they prefer to use.

- **Phase 3: Conditioning**

The purpose of this phase is to draw out into expert's immediate consciousness all relevant knowledge related to the uncertain quantity. Usually, when dealing with a problem, as was mentioned in chapter 4, we have generic (or distributional or base information) and case-specific information. It is important to realise both types of information and combine them adequately. Usually subjects ignore generic information and base their opinions on case-specific information. In order to avoid this situation, the authors propose to ask experts the following question: In your opinion, what would be the answer to this question given by another person with no case-specific information? The answer provided by experts could be taken as a prior distribution that could be combined with the opinion based on only case specific information to get the posterior distribution through the use of Bayes' formula. In this phase experts are also warned about the risk of using data with no predictive capability to make predictions based on them. They propose to use a method suggested by Tversky and Kahneman (1982) to correct such kind of problems, which in fact implies a regression to the mean. These authors propose the following *measure of predictability* $\tau = 2 \cdot \rho - 1$, where ρ is the correlation coefficient (in general a subjective estimate) between predictions and outcomes. Then, if the expert provides an estimate *Y*, when the mean of the prediction is μ_Y the estimate should be corrected towards the value $\mu_Y + \tau \cdot (Y - \mu_Y)$. It must be pointed out that this correction can be used only for absolute values of ρ above 0.5. In any other case (weak correlations described by correlations coefficients between –0.5 and 0.5) the suggestion of the authors is to use the mean (extreme case of regression to the mean).

In this phase, in order to avoid overconfidence, experts are encouraged to think of scenarios that could produce either extremely high or extremely low values of the quantity and think about their likelihood. If needed, experts could be invited to do a calibration exercise.

- **Phase 4: Encoding**

This is the phase where the analyst and the expert together build the distribution function. The techniques used in this protocol are the quantile and the interval techniques, and the combination

of both. Experts are allowed to choose the scale they prefer to make their probabilistic statement (probabilities, odds, log-odds, bets, etc.). Indirect techniques such as the probability wheel are also accepted.

In order to assess joint bivariate distributions, the method suggested is estimating one marginal distribution and then several conditional distributions. In the case of multivariate distributions the method suggested is to estimate firstly all the marginal distributions and then to estimate the median of one variable conditional on a value (excluding the mean) of another variable. This way marginal distributions and measures of correlation are obtained in a simple, though not very precise manner.

Consistency of expert's assessments must be checked from time to time by asking the same or similar questions in different ways. Lack of consistency may be also detected by drawing the estimates provided by the expert on graph paper. In some cases, the first points lie on one line while subsequent points lie on a shifted different one. Usually this means that after a given point in the encoding session the expert started taking into account in the assessment some additional information. Under these circumstances it could be convenient to repeat the encoding session.

- **Phase 5: Verification**

The objective of this phase is to check that the expert agrees with the distribution generated during the encoding phase. The pdf and the cumulative distribution are plotted. In section 5.1.2.2 the advantages of plotting both functions were discussed. Reviewing those plots could highlight some effect not foreseen by the expert that could warrant some further discussion. The final step is to ask the expert if he/she would be willing to bet on his/her estimates. If the expert does not feel comfortable with the distributions generated, some phases of the protocol should be repeated, at least the encoding phase.

- **Phase 6: Aggregation**

This and the next phase are the two phases that were added after the dissolution of the Decision Analysis Group. The original protocol was designed to elicit judgements from only one expert. In many cases it is to gather the judgment from several experts. In that case aggregation would be necessary.

Merkhofer (1987) considers that the first step to correctly aggregate opinions is an efficient exchange of information that would enhance the convergence of the different opinions. This exchange of information should be organised by the analyst based on a variation of the Delphi technique or on the nominal group technique (this makes sure the absence of unacceptable destructive pressures). Both techniques are described in next chapter. Finally the aggregation should be done by mathematical means, preferably using a linear pool (see next chapter), either with equal or with unequal weights. In case that some of the experts do not converge to the opinion of the rest, those distributions could be left apart, to perform sensitivity analysis.

- **Phase 7: Discretising**

In some cases, usually due to demand from the organisation interested in getting the opinion of the experts, it could be convenient to transform the (continuous) distribution obtained into a discrete one. This is done by dividing the range of the variable in several intervals, taking a representative point in each segment and assigning it the probability of the interval. It is convenient to impose some restrictions in this process, as for example preserving the mean and the variance. Miller and Rice (1983) propose a Gaussian quadrature that allows preserving important variable moments.

As a rule of thumb, the estimated time to dedicate to each expert is between half an hour (quite unlikely) and two hours.

5.2.2.- The SNL/NUREG-1150 protocol

The SNL/NUREG-1150 protocol was developed at SNL during the mid-1980's as a collaboration between scientists dedicated to nuclear safety and experts in the field of expert judgement. This protocol was planned as a mean to obtain a lot of information for large-scale risk studies, as for example PSA for NPPs and PA studies for radioactive HLW repositories. The steps of this protocol are the following (USNRC, 1990):

1. Selection of issues
2. Selection of experts
3. Training
4. Presentation of issues
5. Preparation and discussion of analyses
6. Elicitation of experts' opinions
7. Aggregation of results
8. Review
9. Documentation

The different steps are described in detail below.

- **Phase 1: Selection of issues**

In this phase the objective is to identify all issues in the study that could demand the use of expert judgement. The process starts with the development of a first list of issues that could potentially be of interest. At this stage of the process, suggestions from any stakeholder, even from the public, can be accepted. It is preferable to include too many issues in the preliminary list than missing some really important issue. Later on, this original list is screened to keep only those issues that are really relevant to the study. Being relevant means that they must meet two conditions: being uncertain and having a real impact on the results of the study.

Then, each issue must be accurately defined. Usually the starting point for any question to be solved is vague. It is necessary to arrive at a *complete definition* of the parameters whose uncertainty we want to characterise. *Complete definition* of a parameter means the full definition of the parameter, the initial conditions to evaluate it and any other implicit hypothesis under the initial conditions. The final definition should be extremely clear and accurate, with no ambiguity. It should pass the clairvoyant test.

The complete definition of the question includes the way the experts should provide their answers. Since in risk analyses, uncertainty is characterised as Bayesian Probabilities, experts should provide their assessments of uncertainty through these kind of probability. Experts should provide probability distributions, either discrete or continuous.

After the full definition of the question, a list with all relevant sources of information should be produced. Potential decompositions of the parameters could be produced. The list of references to be considered must show the actual state of knowledge in that area, but independence and reliability of the sources should always be kept in mind.

- **Phase 2: Selection of experts**

Three types of experts are considered in this protocol: Subject experts (experts), generalists and normative experts (analysts). In the next paragraphs, a description of the qualities that these experts should have and the process followed to choose them are discussed.

1. Generalists: Generalists supervise the whole process and check the quality of all models used and analyses performed. Usually they come from the organisation interested in obtaining the expert opinions. They should know the details of the study where expert judgement estimations will be used, should have good management capabilities and should also be good at interacting with people. They are the link between the analysts and the experts.
2. Experts: They should deeply know the issue under study and, if possible, be outstanding researchers in that area. The first step in choosing them is to make a list of potential experts. In principle any person or organisation could propose names. A public call to propose experts could be carried out, not excluding self-nominations. The key criterion in choosing an expert is that he/she is a real expert. The selection of an expert is based on: the CV, an interview, and to be acknowledged as expert by his/her colleagues. Another point to consider is his/her willingness to sign up to his assessments. Another point to check is the potential influence of motivational biases. Experts could be asked to provide a list of potential motivational biases that could affect them. It is not frequent to disqualify an expert for this reason. The number of experts employed per issue, attending to some Bayesian arguments (Clemen and Winkler (1985)) about the effect of correlation on the combination of opinions, is usually between 3 and 5. Diversity is the criterion used to obtain independence. Experts chosen should normally work with different information sources, should have a different scientific background (engineers

vs. physicists or chemists), should follow different approaches to the problem (experimental vs. theoretical) and should have different professional experience.

3. Analysts: In addition to having a sound knowledge of probability, statistics and knowledge psychology, they should also have some experience in the use of the protocol. It is also desirable that the analyst has had some contact in the past with scientists and engineers; this helps experts to feel comfortable when giving their assessments. The selection should be based on the CV and on consultation with colleagues.

- **Phase 3: Training**

The objective of this phase is to inform experts about normative aspects of expert judgement elicitation processes. It can be decomposed into the following sub-objectives:

1. Motivate experts to provide rigorous assessments,
2. Remember basic concepts of Probability and Statistics,
3. Training in the assessment of Bayesian probabilities, and
4. Informing experts about basic issues related to knowledge biases.

During the motivation phase the experts must obtain information to point out the importance of the work they are going to do. Firstly, the project team explains to the experts the study frame where their opinions will be used, stressing the part of the study where their opinions are relevant. Secondly, the necessity of expert judgement will be explained, explaining in depth the concept of *Lack of Knowledge Uncertainty,* and how it links to them. Thirdly, the project team will say explicitly that the key issue is not to predict a single value of each parameter under study, but characterising their uncertainty, allowing others to know the actual state of knowledge in that area.

After remembering basic Probability and Statistics concepts, the experts get some training about assessing Bayesian probabilities, which includes: accurate definition of questions to be assessed (making explicit implicit hypotheses, showing well and badly posed questions), decomposition as a way to simplify assessments (use of influence diagrams, event trees and uncertainty propagation techniques) and adequate evaluation of different evidences in order to assess probabilities (use of Bayes' theorem and concepts of independence and reliability of information sources).

The last part of the training session is dedicated to explain Knowledge biases to the experts in order to teach them to provide more reliable opinions, i.e.: representativity, availability and anchor and adjustment. Experts should be informed about the hazard of being overconfident. A calibration exercise could be appropriate. The whole training session should not take more than one morning.

- **Phase 4: Presentation of issues**

This step is done through an interactive session composed of the project team and the experts. The issue at hand is to explain to the experts, in a detailed way, the questions to be assessed and to make a schedule of the activities to be developed by each expert. All the work developed by the project team during the *Definition of the Problem* phase should be used now. The session should start with a presentation given by the generalist about the parameters to be assessed, including all relevant sources of information previously identified. Experts should provide their own view of the problem and the definition of the parameters, pointing out, if needed, further information sources, computations to be made, etc. The result of this session, eventually, would be a refined definition of the parameters under study. Common definitions to all the experts should be agreed.

The second step in this meeting is to study the possible ways to decompose each parameter. The analyst and the generalist should provide a seminal decomposition that should be discussed with the experts. The objective is to help the experts to develop their own decompositions. Decompositions could be quite different from one expert to another one. Expert will have to assess uncertainties of variables at the lowest levels. The analysts will do the appropriate aggregation later. This is the point to introduce propagation of uncertainty concepts to the experts and to inform them about all the potential variety of tools that the analysts could provide them to pre-process and post-process probabilistic runs of computer codes or of the simple decomposition models developed by experts.

- **Phase 5: Preparation and discussion of analyses**

Experts develop their analyses during this phase, according to the schedule agreed in the previous step. By the end of this period, each expert will write a report summarising the main hypothesis and procedures used during his/her work, the conclusions achieved and, if he/she wishes, a preliminary assessment of uncertainties. Whenever needed during this period, the project team should be available to each expert in order to provide statistical support or to resolve any doubt about the parameters to be assessed. At the end of this period, a meeting is organised where experts are encouraged to present their approach to solve the problem. This is a wonderful forum for exchanging information and different points of view. Sometimes, after this meeting, experts review and change their analyses.

- **Phase 6: Elicitation of experts' opinions**

The elicitation of each expert opinion's is individual and is done in a quiet environment, if possible without interruptions. It is convenient to have the presence of an analyst and a generalist, in addition to the expert. In a systematic way, the analyst obtains the opinion of the expert for each parameter, asking for supporting reasoning whenever necessary. The role of the generalist in this session is to provide additional information when needed, to provide general support and to audit the session in order to avoid irregularities (bias induction, etc.). Whenever needed, the analyst could ask questions in a different way to check potential inconsistencies. The session should be recorded (tape recorders, video or extensive hand annotations). This is crucial to deliver a good documentation of the expert judgement exercise.

The techniques used to help the expert when assessing uncertainties are quite standard: quantile assessment for continuous variables and probability estimations for discrete variables (direct or indirect methods); in the case of experts with some skills in probability other techniques like direct parameter assessment or drawings are acceptable.

- **Phase 7: Aggregation of results**

Assessments provided by experts are studied in this phase. The objective is to check that there are not important biases and the logical correctness of their rationale. If biases and logic faults are not present in expert's assessments, the next step is to check if individual opinions may be aggregated to obtain a unique distribution for each parameter.

Before aggregating individual distributions one condition should be checked. It is related to the overlap between distributions of different experts. If the distributions do not overlap, it means essentially that the experts disagree. In that case aggregation should be avoided. Under these circumstances a reconciliation session could help. An analyst should lead the session and should organise it according to the following steps:

1. Exposition of different opinions.
2. Identification of differences.
3. Discussion about the reasons for each original assessment.
4. Discussion about the different sources of information used.
5. Re-elaboration of individual opinions in posterior elicitation sessions or joint assessment (through consensus) of a common distribution, if agreed by experts.

In the case that a consensus distribution is obtained, that is the final step (before documentation). If further elicitation sessions are needed, the consistency of the opinions is checked again and aggregation is done via linear pool with equal weights.

- **Phase 8: Review**

The written analysis of each issue developed by the analyst and the generalist is returned to the experts for review. This review is aimed at avoiding any potential misunderstanding, making sure that the experts' rationale has been correctly summarised by the analyst and the generalist, and actually reflects the experts' opinion.

- **Phase 9: Documentation**

Documentation of the application must be as complete as possible, including results and description of the methods used to obtain them. The contents of the documentation should follow the order of application of the procedure, recording, in each step, *what* has been done, *why* it has been done, *how* it has been done and *who* has done it. In order to achieve this degree of documentation, a schedule of standardised documentation activities should be made for each

phase. It should always be completely clear to the reader what is a result assessed by an expert and what results are the outcomes of an aggregation, sensitivity analysis or any other analysis not provided explicitly by an expert.

5.2.3.- JRC's KEEJAM protocol

This protocol was developed by the Joint research Centre (JRC) of the European Commission (EC) in collaboration with the University of Bolonia. The Knowledge Engineering methodology for Expert Judgement Acquisition and Modelling (KEEJAM) is based on the theory of Knowledge Engineering; see Cojazzi et al. (1997) and Cojazzi and Fogli (2000). This protocol is completely different from any of the protocols previously described and from those described in next chapter.

According to the authors of this protocol, the problem of expert judgement is just a knowledge problem; only an in depth study of the knowledge available to the experts about the problem under study may help solving it. So, the stress must not be put on experts' opinions but on the data, hypotheses and general knowledge sources used by experts.

Analysts must analyse all the knowledge provided by experts to build a self-consistent model. This model should be the best representative of the state of knowledge about the problem under study. It should also be of help to find out the origin of the discrepancies between the experts and problems in the formation of judgements such as lack of accuracy and vagueness, even problems related to the characterisation of uncertainty. The authors admit the possibility of using formalisms different from the theory of probability to address problems, such as the theory of fuzzy sets. The proposed protocol consists of five phases that may be further decomposed in tasks (Cojazzi et al. (1997)):

- **Phase 1: Start**

The problem is analysed in a preliminary way, trying to find out if the problem may be solved with this methodology. The requirements of the model to be developed are defined. The objective is to tailor the methodology to the problem at hand, design the project team and plan the application.

- **Phase 2: Design**

This phase is aimed at defining appropriate techniques for the representation of the types of knowledge and reasoning strategies relevant to the application domain of interest, also including the treatment of the imperfections that may affect knowledge and reasoning.

- **Phase 3: Knowledge acquisition and modelling**

This phase is devoted to acquiring knowledge from the identified knowledge sources (experts, scientific papers and real-world context) and at developing a domain conceptual model that meets the stated requirements.

- **Phase 4: Exploitation and refinement**

This phase is dedicated to exploit the conceptual model developed for carrying out the set of expert judgement tasks.

- **Phase 5: Synthesis and release**

All the results obtained are collected and suitable documentation about the work done is produced.

Probably the major shortcoming of this protocol is the huge effort in terms of work, budget and time that must be devoted to implement it in order to solve complex problems, see Cojazzi and Fogli (2000).

6.- Combination of expert judgement

The objective when developing an expert judgement exercise is to obtain the opinion of the most relevant experts about each parameter or event of interest. Empirical evidence shows that, in general, the aggregated opinion of several experts is better than the individual opinions, though usually some outstanding experts perform better than the group. Because those outstanding experts are sometimes not easily identified and other times they are not available, the common procedure is to count on several experts and to obtain their aggregated opinion.

There are two general types of methods to obtain the aggregated opinion of a group of experts: group methods and mathematical methods. Group methods allow the interaction among experts to arrive at a consensus opinion, while mathematical methods consist of different mathematical techniques to combine individual opinions and produce a joint opinion consistent with the individual ones. Among the group methods, the most used and the best-known are the total interaction group method, the nominal group method and the Delphi method. Mathematical methods may be classified in two groups: pools (linear and log-linear) and Bayesian methods. No consensus exists about which methods produce better results. Each method shows advantages and drawbacks. Depending on the type of task and the objectives of the exercise, some methods could be more appropriate than others.

6.1.- General characteristics of expert judgement combination

Before getting into the technical details of the different methods to aggregate the opinions of different experts, it seems convenient to consider some aspects related to the possible causes of discrepancies between experts and the aggregation method to be used.

As a general rule, whenever discrepancies between experts occur, it is convenient to study them in order to identify their origin and possible consequences. In fact, discrepancies become important sources of information. According to Roberds (1992) the origin of different individual assessments may be classified as

1. Disagreement on the assumptions or definitions that underlie assessments.
2. Failure to overcome assessment errors and biases.
3. Judgements based on different information sources.
4. Disagreements on how to interpret available information.
5. Different opinions or beliefs about the quantity of interest.

The analyst must take into account these ideas carefully in order to detect the existence of non-admissible differences, meeting the experts again if needed to check some of their opinions. As an example, contacting an expert again could be justified if the analyst considers that that expert did not take into account some relevant piece of information. Identifying some lack of accuracy

in the formulation of the questions to be assessed demands a new and accurate formulation and the corresponding re-assessment by the experts.

If, eventually, the differences among experts are admissible, then it is convenient to check how those legitimate differences may affect the results of the study where expert opinions are going to be used: a sensitivity analysis would be necessary. If this analysis shows that the differences among the opinions of different experts does not produce a big difference in the results, then the aggregation of the different opinions is justified and this aggregation does not produce any significant loss of information. But, if the differences do actually have a high impact on the results of the study, considerable care should be taken. Two main options are available. The first one is to aggregate only the opinions that overlap substantially, keeping the outliers for performing sensitivity analysis. The second one, acceptable when disagreements are very large, consists of not performing any aggregation, accepting that the problem is very much affected by uncertainty.

When disagreements do exist, there are three possible results of the attempt to obtain one single group opinion:

1. *Convergence*. The different opinions of the experts evolve, through information sharing and discussion, towards a unique opinion that represents the final group opinion, and all the members of the group accept that opinion.
2. *Consensus*. The group arrives at a unique opinion that does not represent the view of all the members. The consensus may be forced (not all experts accept the final group opinion) or reached by agreement (all experts accept the final group opinion; some of them accept not to include some of their opinions).
3. *Disagreement*. Huge differences in their opinions avoid any kind of consensus.

Convergence is the most defensible situation when the common opinion is subject to great scrutiny, as in a regulatory review or in a peer review process, though it is usually the most difficult to achieve. Consensus by agreement is a little less defensible and is also difficult to obtain. Forced consensus is relatively easy to achieve but it is hardly defensible.

Group interaction allows experts share information, discuss and combine opinions. Sharing information and inferential processes is a good strategy to reconcile differences. Aggregation by interaction may be a very effective process when different assumptions are the origin of the disagreement. In these cases, discussion helps to uncover such different hypotheses and reach either convergence or consensus. The most important threat to group interaction is the existence of destructive interpersonal relations among experts. Another shortcoming of this procedure to reconcile differences is the large organisational effort that these meetings involve, not to speak about the problems related to the project budget. Morgan and Henrion (1990) think that this type of method is of little use in the scientific and technological fields for most experts are aware of their colleague's opinions (scientific papers, interaction in conferences and joint past projects, etc.), and the room for changing opinions is small.

Mathematical aggregation methods show some advantages: it is relatively easy to use if the analysts have the required mathematical background, it is relatively easy to perform uncertainty and sensitivity analysis and destructive interpersonal relations among experts are avoided. These methods are very much recommended when disagreements are not large and cannot converge further. The main drawbacks are: consensus achieved by means of these methods is certainly forced and, in many cases they involve the use of subjective judgements made by analysts to assign (different) weights to the experts.

Empirical evidence does not provide a definitive support to either type of methods. Both strategies show advantages and drawbacks. In fact, the advantages shown by one type of methods are at the same time the drawbacks of the other type and vice versa: mathematical aggregation avoids destructive interpersonal relations among experts but, at the same time, it precludes sharing information and knowledge, what certainly is undesirable in a process where obtaining information and knowledge is key to obtaining the right opinions. So, the use of one of either type of methods depends very much on the problem at hand, the magnitude of the discrepancies and their origin. Moreover, group methods may work to address any type of uncertainty, while mathematical methods are only suited for combining opinions about events and parameters. Another option is using both types of methods together. In this case, group interaction can be used to guarantee an efficient share of information, hypotheses and rationale, but the final aggregation of individual opinions may remain mathematical.

6.2.- Group combination

The formation and modification of group opinions have been deeply studied by Social Psychology. Researchers in this area of science have established that, on average, the quality of group opinions is higher than the mean opinion of individuals. Nevertheless, regarding the most complex tasks, the group does not usually perform as well as the most accurate individual in the group. Moreover, two phenomena that appear in groups may introduce important biases in their common opinions: dysfunctions and preference shift. The most important dysfunctions that may appear in a group are the following:

1. An effect of central tendency that takes the group to follow a limited set of lines of thinking.
2. An effect of auto-weight. Each member of the group participates in the group debate and tries to influence on the group opinion proportionally to his/her opinion about his/her own competence in the field.
3. An effect of hidden agenda, that makes some experts not give their real opinions openly to the rest of the group.
4. Group pressure on some members to reach a consensus.
5. Influence of the strongest personalities.

The preference shift consists in a change in the opinion of subjects produced by psychological factors not related to the task under study. There is some empirical evidence that in some group

discussions experts shift their original opinions towards other opinions that involve a higher risk. This phenomenon is due to the fact that group discussion usually reduces the level of anxiety of individuals about the possible adverse consequences because the responsibility is considered to be shared by all the group members.

The identification of these problems of group behaviour has provided some guidance to design methods to obtain group opinions. The aim is to enhance the exchange of information, rationale and opinions, avoiding as much as possible at the same time the introduction of group biases. Three methods have been selected for discussion in this chapter: the total interaction group, which allows open and unrestricted discussions among experts, the Delphi method, which allows interaction among experts via exchange of written opinions, without direct interaction, and the nominal group method, which allows direct interaction among experts moderated by an analyst.

6.2.1.- Total interaction group

The aim of this method is to reach a group opinion about an issue, either via convergence or via consensus and accepting non-restricted interaction among experts. The main drawback of this method is the possible concurrence of group biases. Other problems that can also arise in an application of this method, for example a poor quantification of the uncertainty or the existence of unknown implicit assumptions, may be a result of the degree of formality of the method (more formal applications producing better quality results in general). If, before the discussions, the problem has not been clearly and accurately defined, or if experts have not provided a first individual opinion to be used as a starting point for discussion, the quantification of uncertainty can be poor.

Some authors think that this technique may be of interest to deal with uncertainty problems in a qualitative manner, where oral communication may be very important to exchange information.

6.2.2.- The Delphi method

The Delphi method was developed in the late 1940s and early 1950s by the Rand Corporation for the USA army as a tool to improve the quality of the decisions made via consensus in the military area. In the middle of 1960s and early 1970s it found a wide variety of applications (Dalkey (1968) and Brown et al. (1969)). Its use has clearly decayed since the 1980s onwards. The method was mainly applied to technology forecasting, but also to different types of policy analyses.

The method has three main features: 1) anonymous response, 2) iteration and controlled feedback and 3) statistical group response. The method may be structured in the following steps:

1. A questionnaire is developed and sent to the experts previously selected.

2. Each expert gives his/her answers to the questions in an independent and anonymous way.
3. The responses of each expert are analysed by the monitoring team. The lower 25% and the upper 25 % of responses are excluded.
4. The set of remaining responses is then sent back to experts and they are asked if they wish to revise the initial predictions.
5. The process is iterated until experts reach a certain consensus.

The Delphi method has undergone many variations. One of the most important was letting the experts indicate their own expertise for each question (for example rating their expertise on a scale of 1 to 7). This variation was supposed to improve accuracy, since only opinions of experts with "higher" expertise were used to determine the distribution of opinions for that item. This approach was challenged when it was found that self-rating of participants did not coincide with "objective expertise". Moreover, it has been found that women consistently rate themselves lower than men.

6.2.3.- The nominal group method

This method profits from the advantages of the face-to-face discussion, trying simultaneously to avoid group biases. This is achieved by means of an analyst that moderates the group discussions. The controlled face-to-face interaction is more dynamic and avoids experts tiring because of having to provide written arguments in support of their opinions, as can happen when applying the Delphi method. Each expert always provides an individual first assessment before the discussion meeting. Those assessments are the starting point for the discussion. Roberds (1992) describes this formal process as a six-step method:

1. Motivation
2. Identification of differences
3. Discussion about the reasons for each initial assessment
4. Discussion about the information sources used
5. Re-elaboration of individual assessments
6. Reconciliation of differences

The role of the analyst is, in addition to avoiding group biases, to interact with experts in order to reach consensus about definitions, hypotheses, information sources used, data interpretation, etc. In case a consensus is not achieved, mathematical aggregation is kept as an option.

6.2.4.- The protocol used by Nirex and the NDA in the UK

In the mid 1980's the UK Department of Environment studied an expert judgement group consensus methodology. The study was successful and it showed that it was feasible to apply it to obtain relevant information from a group of experts. Based on the experience acquired, the

UK DoE used that methodology to elicit several pdf's within the European project PACOMA (Dalrymple and Phillips (1987)). This work was much criticised in a peer review process (Zimmerman et al. (1991)), especially because it used a group consensus methodology instead of eliciting individual distributions and combining them via a mathematical method. Nevertheless, UK DoE rejected this criticism on the grounds of defensibility of the consensus distributions. Since consensus was not forced, each expert would be willing to defend the common distribution. UK DoE believed that the mathematical combination is not as defensible because it involves elements beyond the pure expert opinions.

Nirex adopted and slightly modified this method and applied it for the first time in 1991 and 1992 (Phillips and Wisbey (1993)). In 2007, Nirex was incorporated into the Nuclear Decommissioning Authority in the UK. This method remains NDA's method of choice for the elicitation of uncertainties for key safety significant parameters in a PA. The method consists of the following phases (Nirex (2006)):

1. **Establishing roles**

Four different roles are defined for the elicitation session

- *Customer*: he/she is a representative of the organisation interested in obtaining the pdf's. That person knows and brings an understanding on how the parameters of interest will be used in the PA. He/she actively participates in the accurate definition of the parameters to be studied and also participates in all meetings to answer any question that could arise about the parameters, but should never influence the elicitation.
- *Facilitator/analyst*: his/her role is to guide the elicitation sessions ensuring that all views are considered and group biases are avoided.
- *Experts*: these persons are those that actually solve the problem. The requirements imposed on them are the same as in all other formal methods.
- *Observers*: their presence is not mandatory. They can contribute to the general discussion but they don't provide any opinion.

2. **Determining the scope of the project**

The customer, together with the facilitator, discusses the importance and role of the parameters in the PA model. They have to make sure that the definitions of the parameters are understood and agreed. The customer may also define the degree of detail needed in the pdf's to be assessed: the need of only upper and lower limits, the use of specific distribution (normal, triangular, etc.) or the imposition of no restriction on the shape of the distribution, allowing the experts to fully determine it.

3. **Preparation of the meeting**

The customer and the facilitator decide about the number and identity of the experts that will participate, possibly in consultation with experts known to the customer. They also decide whether observers will or will not be invited. The customer prepares a data pack to define the extent of the search for appropriate data. The data pack is distributed to the facilitator and to the experts at least two weeks in advance of the elicitation session.

4. **Meeting protocol**

The protocol used to perform the elicitation session is the SRI protocol, see chapter 4 in this document. The only innovation with respect to the SRI protocol is the inclusion of the customer as a person that provides information at demand of the experts and that resolves any doubt about the definition of the elicited parameters.

5. **Reporting the results of the meeting**

During and after the meeting the customer and the facilitator prepare a report containing the results of the elicitation session and the rationale that supports such results.

6. **Review of elicited distributions**

The elicitation report, prior to its final publication, is distributed for peer review to experts that have not participated in the elicitation. The review comments cannot alter the results of the elicitation but under exceptional circumstances, as for example omission of relevant sources of information, can lead to a re-elicitation. After this peer review the elicitation report is finally published.

As a final remark about group methods, no conclusive experimental results are available, but the expected benefits of avoiding important group biases make the Delphi method, the nominal group method and other methods like the method used by Nirex and NDA more attractive to analysts.

6.3.- Mathematical aggregation

As was mentioned at the beginning of this chapter, there are two general methods to address the problem of aggregating mathematically the opinions of different experts: pools (linear and log-linear) and Bayesian methods. In both cases a person or group of person acts as a Decision-Maker (DM), who has his/her/its own opinions about the issues under study and about the quality of the experts. In the linear (log-linear) pool the DM has to attribute to the distribution provided by each expert, a relative importance according to how much credibility the DM thinks each expert and the information he/she uses, deserves. When applying Bayesian methods, the DM has his/her own (a priori) opinion about the issue and updates it by means of Bayes formula, using the experts' opinions as empirical evidence. Linear pools are more easily applicable, but no

attempt has been made to incorporate the effect of possible dependence between experts on the assessment of their relative weights, while Bayesian methods are able to deal with this problem.

6.3.1.- The linear pool

Suppose *m* experts have provided their respective distributions for a given uncertain parameter, θ, expert *j* provides distribution $f_j(\theta)$ and so on; suppose each expert is assigned a weight ω_j Contained in the interval [0,1], subject to the restriction

$$\sum_{i=1}^{m} \omega_i = 1 \quad . \tag{6.1}$$

Then, the expression

$$f(\theta) = \sum_{i=1}^{m} \omega_i f_i(\theta) \tag{6.2}$$

is the linear pool of the individual distributions. Under the same conditions, the log-linear pool is defined as

$$\log(f(\theta)) = \sum_{i=1}^{m} \omega_i \log(f_i(\theta)) \tag{6.3}$$

The only problem to apply the linear pool and the log-linear pool is the determination of the set of weights. It is convenient to keep in mind that the weight assigned to each expert could vary from one parameter to the next one. Three methods are available in the literature to assess weights to experts: equal weights, Saaty's method and Cooke's classical method.

6.3.1.1.- Equal weights

This is the case when we have *n* experts and a weight 1/n is assigned to each expert. Equal weights is a defensible strategy when there is no reason to think that any expert performs better than any other one, but it is not so defensible when differences in the quality of experts are evident, neither when strong dependences among experts are detected (dependent information is less valuable than independent information from an inferential point of view). This is a widely used strategy.

6.3.1.2.- Saaty's method

Saaty (1988) developed the Analytical Hierarchy Process (AHP) as a support tool to be used in the framework of multi-attribute decision analysis to establish a hierarchy and sort attributes according to the DM's opinions and preferences. AHP is easily adaptable to the task of assigning weights to the opinions of different experts.

Saaty's method arises from the explicit acknowledgement that human beings as DM's have many difficulties to simultaneously compare many factors (experts). Nevertheless, human beings are efficient at comparing two possible alternatives, as for example when they are asked to rank them by means of using expressions such as better, worse or equal. Following this method, experts are assigned weights according to the subjective quality that the DM attributes to each expert. The main advantage of this technique is that it is easy to learn and to apply.

Consider a DM that has obtained the opinion of *n* experts about a given issue. Then, he/she is asked which expert is best, *j* or *k*. In total, he/she will have to answer this question n(n-1)/2 times, the number of pairs of experts that we make take out of *n* individual experts. Three real values a>1, 1 and 1/a. are associated respectively to the adjectives better, equal and worse. Saaty recommends to take $a=e$ when the choices are the above mentioned better-equal-worse. The values obtained from the n(n-1)/2 comparisons are used to fill in a $n \times n$ square matrix **A** whose element a_{jk} is the value assigned to the relation between experts *j* and *k*. The diagonal elements are all equal to 1. Such a matrix for 4 experts could be as follows:

$$\begin{pmatrix} 1 & 2.72 & 2.72 & 2.72 \\ 0.37 & 1 & 1 & 0.37 \\ 0.37 & 1 & 1 & 0.37 \\ 0.37 & 2.72 & 2.72 & 1 \end{pmatrix} \qquad (6.4)$$

which represents the case when expert 1 is the best, experts 2 and 3 are equally good and worse than expert 4. In order to get the weights for the different experts Saaty proceeds in the following way: 1) compute the eigen values of matrix **A** and the eigenvector, $\Omega^T = (\Omega_1,...,\Omega_n)$, associated to the largest eigen value, λ_{max}, of matrix **A** (remember that $A\Omega = \lambda_{max}\Omega$). The weight associated to each expert is the ration between each component of that eigenvector and the addition of all its components ($\omega_i = \Omega_i \Big/ \sum_{j=1}^{n} \Omega_j$). In the example represented by matrix (6.4), the corresponding weights obtained for experts 1, 2, 3 and 4 are respectively 0.46, 0.13, 0.13 and 0.28.

When the number of experts is small it is easy to check whether the DM has made any consistency mistake in the paired comparisons (lack of transitivity of preferences). When the number of experts is large, checking coherence by inspection of matrix **A** may be cumbersome. For those cases the author developed a consistency index that allows detecting lack of consistency in the preferences of the DM, see Saaty (1988).

6.3.1.3.- Cooke's classical method

This method was designed to avoid arbitrariness in the assignation of weights to experts obtained via purely subjective opinions of the DM. Cooke proposes to base the assignation of weights on the precision of the expert. According to Cooke (1991), an expert is precise if he/she is well calibrated and his/her opinions are informative. As it was mentioned in chapter 4, an expert is well calibrated when his/her assessed probabilities agree with actual observed frequencies. The informative character of a pdf is related to its dispersion, the less disperse it is, the more

informative. The more disperse a distribution is, the less it may be used with predictive purposes. So, an expert is good, and his opinions deserve being highly weighted, when they are calibrated and are informative. Cooke (1991) has developed a scoring rule that measures simultaneously both properties (calibration and informativeness). The main problem with Cooke's method is the design of the right set of parameters, called *seed variables*, used to assign weights. These parameters must be in the domain of expertise of the experts that participate in the expert judgement exercise and their actual values must be known to the analysts and unknown to the experts. See Cooke (1991) for more details.

6.3.2.- Bayesian combination of expert judgement

When a DM adopts a fully Bayesian approach, Bayes' formula is an adequate tool to combine experts' opinions in order to update his/her own state of knowledge about the issue under study, Let θ be the parameter whose uncertainty the DM wants to characterise. The a priori state of knowledge of the DM about that parameter is represented by $P(\theta|H)$, the prior distribution of the DM for parameter θ conditional on all his/her knowledge, H. Experts' opinions is equivalent to new information H' given as a multivariate distribution that indicates what values of θ experts consider more likely. The DM will combine both pieces of information by means of Bayes' formula

$$P(\theta / H, H') \propto P(H' / H, \theta) \cdot P(\theta / H) \quad , \tag{6.5}$$

where the left hand side is the a posteriori DM's distribution for θ after collecting experts' opinions and the first factor on the right hand side is the likelihood of experts' opinions in the opinion of the DM. The likelihood is a key element in Bayes' formula used by the DM to model the expert. Lindley (1988) is one of the best available papers about Bayesian combination of experts' opinions.

7.- Conclusions

Expert judgement is a technical discipline, between science and art, which started its development shortly after the end of World War II. Since then a lot of research has been done about the way people make judgements, the problems they may encounter and the way to counteract them. After the pioneering Delphi method, several other protocols have been developed to make sure that subjects' opinions are obtained as free of biases as possible. The need to incorporate explicitly uncertainties in risk analyses of complex industrial facilities, and specifically the need to do this for the PSA of NPPs and for the PA of radioactive HLW repositories, triggered the development of specific protocols in the nuclear field, such as the protocols SNL/NUREG-1150, KEEJAAM and the protocol used by Nirex and the NDA in the UK. In this report the authors have provided an overview of all issues related to expert judgement and protocols to obtain expert judgement in a formal and structured way. This report is expected to be used as training material for experts that are going to participate in formal processes to get their opinions about technical and scientific matters.

References

R.E. Barlow (1988). Using Influence Diagrams. In 'Proceedings of the CII Course of the International School of Physics "Enrico Fermi" on Accelerated Life Testing and Expert Opinions on Reliability'. North Holland.

T. Bayes (1958). Essay towards solving a problem in the doctrine of chances. Biometrica, Vol. 45, pp 293-315.

R.L. Beach and T.S. Scopp (1968). *Intuitive statistical inferences about variances.* Organizational Behavior and Human Performance, Vol. 3, pp 109-123.

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

E.J. Bonano, S.C. Hora, R.L. Keeney and D. von Winterfeldt (1990). Elicitation and use of expert judgement in performance assessment for high-level radioactive waste repositories. Sandia National Laboratories, SAND89-1821 (NUREG/CR-5411).

K. Chaloner and G.T. Duncan (1987). *Some Properties of the Dirichlet-Multinomial Distributionand its Use in Prior Elicitation.* Communications in Statistics, Theory and Methods, Vol. 16 (2), pp. 511-523.

R.T. Clemen and R.L. Winkler (1985). *Limits for the Precision and Value of Information from Dependent Sources.* Operations Research, Vol. 33, pp. 427-442.

G. Cojazzi, G. Guida, L. Pinola, R. Sardella and P. Baroni (1987). KEEJAM: a Knowledge Engineering Methodology for Expert Judgement Acquisition and modelling in Probabilistic Safety Assessment. In 'Advances in safety and reliability', Vol. 1, pp. 199-206. C. Guedes Soares (Editor), Pergamon.

G. Cojazzi and G. Fogli (2000). Benchmark exercise on Expert Judgement Techniques in PSA Level 2, Extended Final Report. EUR 19739 EN.

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

N.C. Dalkey (1969). An Experimental Study of Group Opinion. Rand Corporation Report RM-5888-PR.

G. Dalrymple and L.D. Phillips (1987). Using a Structured Approach to the Acquisition of Probabilistic Data from Expert Opinion. CAP Scientific Limited. London. Report No. 3409/TR.2.

B. De Finetti (1964). Foresight: its Logical Laws, its Subjective Sources. In 'Studies in Subjective Probability'. H.E. Kyburg and H.E. Smokler Eds. Wiley.

B. De Finetti (1974). Theory of Probability. John Wiley & Sons.

M.H. De Groot (1988). Modern Aspects on Probability and Utility. In 'Proceedings of the CII Course of the International School of Physics "Enrico Fermi" on Accelerated Life Testing and Expert Opinions on Reliability'. North Holland.

DOE (1986). A multiattribute utility analysis of sites nominated for characterization for the first radioactive waste repository – A decision aiding methodology. Office of civilian radioactive waste management. DOE/RQ-0074.

H.J. Einhorn and R.M. Hogarth (1978). *Confidence in Judgement: Persistence of the Illusion of Validity.* Psychological Review, Vol. 85, No. 5.

P.H. Garthwaite, J.B. Kadane and A. O'Hagan. *Statistical Methods for Eliciting Probability distributions.* Journal of the American statistical Association, Vol. 100, No. 470, pp. 680-700.

D.V. Gokhale and S.J. Press (1982). *Assessment of a prior Distribution for the Correlation Coefficient in a Bivariate Normal distribution.* Journal of the Royal statistical Society, series A (1982), 145, Part 2, pp. 237-249.

J.M. Hampton, P.G. Moore and H. Thomas (1973). *Subjective Probability and its Measurement.* Journal of the Royal statistical Society, series A (1973), 136, Part 1, pp. 21-42.

R.M. Hogarth (1975). *Cognitive Processes and the assessment of Subjective Probability Distributions.* JASA, Vol. 70, No. 350, 1975: 271-294.

R.M. Hogarth (1980). Judgement and Choice: the Psychology of Decisions. John Wiley & Sons.

J.B. Kadane, J.M. Dickey, R.L. Winkler, W.S. Smith and S.C. Peters (1978). Interactive Elicitation for a Normal Linear Model. Technical Report 150, Department of Statistics, Carnegie Mellon University, Pittsburgh.

S. Kaplan and B.J. Garrick (1980). On the quantitative definition of Risk. Risk Analysis, Vol. 1, No. 1, pp. 11-27.

A. Kolmogorov (1956). Fundations of the theory of probability. Chelsea Publishing Co.

J.P. Kotra, M.P. Lee, N.A. Eisenberg and A.R. de Wispelare (1996). Branch technical position on the use of expert elicitation in the High-Level Radioactive Waste program. USNRC, NUREG-1563.

P.S. Laplace (1951). A Philosophical essay on probabilities. Dover.

S. Lichtenstein, P. Slovic, B. Fischhoff, M. Layman and B. Combs (1978) . *Judged Frequency of Lethal Events*. Journal of Experimental Psychology: Human Learning and Memory, 4: 551-578.

D.V. Lindley (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge University Press.

D.V. Lindley (1988). Introduction to 'Proceedings of the CII Course of the International School of Physics "Enrico Fermi" on Accelerated Life Testing and Expert Opinions on Reliability'. North Holland.

D.V. Lindley, A. Tversky and R.V. Brown (1979). *On the Reconciliation of Probability Assessments.* Journal of the Royal statistical Society, series A (1979), 142, Part 2, pp. 146-180.

H.W. Lewis, R.J. Budnitz, H.J.C. Kouts, F. Hippel, W. Lowenstein and F. Zachariasen (1975). Risk assessment group report to the US Nuclear Regulatory Commission. USNRC.

J.E. Matheson and R.L. Winkler (1975). *Scoring rules for probability distributions*. Management Science, Vol. 22, No. 10, pp 1087-1096.

M.W. Merkhofer (1987). *Quantifying Judgemental Uncertainty: Methodology, Experiences and Insights.* IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-17, No.5, pp 741-752.

A.C. Miller and T.R. Rice (1983). *Discrete Approximations of Probability Distributions.* Management Science, Vol. 29, pp. 352-362, Mar. 1983.

M.G. Morgan and M. Henrion (1990). Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

A. Mosleh, G. Apostolakis and V.M. Bier (1988). *A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment.* Reliability Engineering and System Series, 20, 1988: 63-85.

T.M. Mullin (1986). Understanding and Supporting the Process of Probabilistic Estimation. Ph. D. Dissertation. Carnegie Mellon University, Pittsburgh.

A.H. Murphy and R.L. Winkler (1984). *Probability forecasting in meteorology.* JASA, Vol. 79, No. 387, pp 489-500.
Nirex (2006). A Procedure for Data Elicitation in Support of Performance Assessments. Nirex report No. N/132.
C.R. Peterson and A. Miller (1964). *Mode, median and mean as optimal strategies.* Journal of Experimental Psychology, Vol. 68, pp 363-367.
L.D. Phillips and S.J. Wisbey (1993). The Elicitation of Judgemental Probability Distributions from Groups of Experts: a Description of the Methodology and Records of Seven Formal Elicitation Sessions Held in 1991 and 1992. Nirex Report NSS/B101.
F.P. Ramsey (1926). Truth and Probability. In 'The foundations of Mathematics and other Logical essays'. Kegan.
Rasmussen et al. (1975). Reactor Safety Study – An assessment of accident risks in US commercial Nuclear Power Plants. USNRC, WASH-1400 (NUREG-75/014).
W.J. Roberds (1992). Methods for Developing Defensible Subjective Probability Assessments. Golder Associates, Report 925-2036. September 1992.
J. Rohrbaugh (1979). *Improving the Quality of Group Judgement: Social Judgement and the Delphi Technique.* Organizational Behavior and Human Performance, 24, 1979: 73-92.
T.L. Saaty (1988). Mathematical Methods of Operations Research. Dover.
L.J. Savage (1954). The Foundations of Statistics. Wiley.
L.J. Savage (1954). The Foundations of Statistical Inference. Methuen.
J. Spencer (1961). *Estimating Averages.* Ergonomics, Vol. 4, pp 317-328.
A.Tversky, D. Kahneman (1974). *Judgement under Uncertainty: Heuristics and Biases*. Science, New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131.
A.Tversky, D. Kahneman (1982). Evidential Impact of Base Rates. In 'Judgement under Uncertainty: Heuristics and Biases', D. Kahneman, P. Slovic and A. Tversky (Editors). Cambridge University Press.
USNRC (1990). Severe accident risks: An Assessment for Five U.S. Nuclear Power Plants. NUREG-1150.
USNRC (1991). Staff's Approach for Dealing with Uncertainties in Implementing the EPA HLW Standards. SECY-91-242. August 6, 1991.
USNRC (1994). A Review of NRC Staff Uses of Probabilistic Risk Assessment. NUREG-1489.
USNRC (1996). Branch Technical Position on the Use of Expert Elicitation in the High-Level Radioactive Waste Program. NUREG-1563.
R. von Mises (1957). Probability, Statistics and Truth. MacMillan Publishing Co.
R.L. Winkler (1967). *The Assessment of Prior Distributions in Bayesian Analysis.* JASA, Vol. 62, No. 319, pp 776-800.
R.L. Winkler and A.H. Murphy (1968). *'Good' Probability Assessors.* Journal of Applied Meteorology, 7: 751-758.
D.A. Zimmerman, E.J. Bonano, P.A. Davies, C.P. Harlan and M.S.Y. Chu (1991). Peer Review of the UK DoE Dry Run 3 Exercise. Sandia National laboratories.