2009

Expert Judgment in Risk Analysis

Stephen C. Hora

University of Hawaii at Hilo, hora@usc.edu

Follow this and additional works at: http://research.create.usc.edu/nonpublished_reports

Recommended Citation
http://research.create.usc.edu/nonpublished_reports/120

This Article is brought to you for free and open access by CREATE Research Archive. It has been accepted for inclusion in Non-published Research Reports by an authorized administrator of CREATE Research Archive. For more information, please contact gribben@usc.edu.
Expert Judgment

Judgment involves the weighing of available evidence and reaching a balanced conclusion from that evidence. We bring in experts to provide these judgments because they have developed the mental tools needed to make sound evaluations. These mental tools include knowledge of what evidence can be brought to bear on the question, the abilities to weigh the validity of various pieces of evidence and to interpret the relative importance of various facts or assertions, and to craft a view from an ensemble of information that may be inherently limited or self- conflicted. In risk analysis these judgments nearly always entail uncertainty so that the judgments are not definitive but reflect what we know and what we know we do not know.

The natural language of uncertainty is probability. It provides a precise way to encode knowledge that is inherently imprecise. It puts judgments in a form where they can be manipulated mathematically, and thereby be integrated with other pieces of information and used in models to assess risks. These judgments are essential in many analyses of risk. For example, see [1–5].

When we set about acquiring expert judgments for an analysis, there are a number of decisions that must be made about how to proceed. These include the following:

• selecting the issues to be addressed by the experts
• selecting the experts
• organizing the effort
• choosing a method for combining multiple judgments, if needed.

Posing Questions to the Experts

The first stage of developing an expert judgment process is to determine the objectives and desired products. Judgments can be made about a number of different things. Some judgments are about facts, while others are about values. Roughly speaking, a fact is something that can be verified unambiguously, while a value is a measure of the desirability of something.

For example, you might believe that it is better to expend tax dollars to fight HIV abroad than it is to improve the educational level of impoverished students at home. This is a value judgment. You might even be able to express in quantitative terms your affinity for one use of tax dollars versus another. But another person may hold very different views and neither of you can be said to unequivocally correct. It is a matter of preference, not of fact. In contrast, you and another person may hold different views about the rate of HIV infection. In principle, it is possible to determine this rate (through medical testing and statistical analysis) and thus the judgments are about a fact. This article is limited to judgments about facts, while the discussions in the articles on Utility Functions and Preference Functions focus on values.

Judgments about facts can be further classified. Focusing on judgments often made in risk analysis, we find that judgments can be made about

• the occurrence of future events
• the values of parameters
• the appropriateness of competing models in their ability to reflect reality.

Not all questions, however, are of equal consequence in quantifying risk. There will normally be a few major questions that drive the uncertainty about the risk. These questions are candidates for a more structured expert judgment activity. Other issues – those that play a minor role – can often be treated less formally or through sensitivity analysis, saving the resources for the more important issues. A sensitivity analysis using initial estimates of probabilities and probability distributions is often performed after an initial risk model has been structured. The sensitivity analysis identifies those questions deserving of a more penetrating study.

However, not all issues lend themselves to quantification through expert judgment. In addition to being important contributors to uncertainty and risk, an issue that is a candidate for expert judgment analysis should satisfy the following conditions:

• It should be resolvable in that given sufficient time and/or resources, one could conceivably learn whether the event has occurred or learn the value of the quantity in question. Hence, the issue concerns a fact or set of facts.
• It should have a basis upon which judgments can be made and can be justified.

The requirement of resolvability means that the event or quantity is knowable and physically
measurable. We consider a counterexample. In a study of risk from a radioactive plume following a power plant failure, a simple Gaussian dispersion model of the form \( y = ax^b \) was employed [1]. In this model, \( a \) and \( b \) are simply parameters that give a good fit to the relation between \( x \), downwind distance, and \( y \), the horizontal width of the plume. But not all experts subscribe to this model. More complex alternatives have been proposed with different types of parameters. Asking an expert to provide judgments about \( a \) and \( b \) violates the first principle above. One cannot verify if the judgments are correct, experts may disagree on the definition of \( a \) and \( b \), and experts who do not embrace the simple model will find the parameters not meaningful. It is very difficult to provide a value for something you do not believe exists.

The second requirement is that there is some knowledge that can be brought to bear on the event or quantity. For many issues, there is no directly applicable data so that data from analogs, models using social, medical or physical principles, etc., may form the basis for the judgments. If the basis for judgments is incomplete or sketchy, the experts should reflect this by expressing greater uncertainty in their judgments.

Once issues have been identified, it is necessary to develop a statement that presents the issue to the experts in a manner that will not color the experts’ responses. This is called framing the issue. Part of framing is creating an unbiased presentation that is free of preconceived notions, political overtones, and discussions of consequences that might affect the response. Framing also provides a background for the question. Sometimes there are choices about whether certain conditions should be included or withheld from the analysis and whether the experts are to integrate the uncertainty about the conditions into their responses. For example, in a study of dry deposition of radioactivity, the experts were told that the deposition surface was northern European grassland, but they were not told the length of the grass which is thought to be an important determinant of the rate of deposition [1]. Instead, the experts were asked to treat the length of grass as an unknown and to incorporate any uncertainty that they might have into their responses. The experts should be informed about those factors that are considered to be known, those that are constrained in value, those that are uncertain, and, perhaps, those that should be excluded from their analyses.

Finally, once an issue has been framed and put in the form of statement to be submitted to the experts, it should be tested. The best way to do this testing is through a dry-run, with stand-in experts who have not been participants in the framing process. Although this seems like a lot of extra work, experience has shown that getting the issue right is both critical and difficult [2]. All too often, the expert’s understanding of the question differs from what was intended by the analyst who drafted the question. It is also possible that the question being asked appears to be resolvable to the person who framed the question, but not to the expert who must respond.

Selecting the Experts

The identification of experts requires that one develop some criteria by which expertise can be measured. Generally, an expert is one who “has or is alleged to have superior knowledge about data, models and rules in a specific area or field” [3]. But measuring against this definition requires one to look at indicators of knowledge rather than knowledge per se. The following list contains such indicators:

- research in the area as identified by publications and grants
- citations of work
- degrees, awards, or other types of recognition
- recommendations and nominations from respected bodies and persons
- positions held
- membership or appointment to review boards, commissions, etc.

In addition to the above indicators, experts may need to meet some additional requirements. The expert should be free from motivational biases caused by economic, political, or other interest in the decision. The choice of whether to use internal or external experts often hinges on the appearance of motivations biases. Potential experts who are already on a project team may be much easier to engage in an expert judgment process, but questions about the independence of their judgments from project goals may be raised. Experts should be willing to participate and they should be accountable for their judgments [6]. This means that they should be willing to have their
names associated with their specific responses. At times, physical proximity or availability will be an important consideration.

How the experts are to be organized also impacts the selection. Often, when more than one expert is used, the experts will be redundant of one another, meaning that they will perform the same tasks. In such a case, one should attempt to select experts with differing backgrounds, responsibilities, fields of study, etc., so as to gain a better appreciation of the differences among beliefs. In other instances, the experts will be complementary, each bringing unique expertise to the question. Here, they act more like a team and should be selected to cover the disciplines needed.

Some analyses undergo extreme scrutiny because of the public risks involved. This is certainly the case with radioactive waste disposal or purity of the blood supply. In such instances, the process for selecting (and excluding) experts should be transparent and well documented. In addition to written criteria, it may be necessary to isolate the project staff from the selection process. This can be accomplished by appointing an independent selection committee to seek nominations and make recommendations to the staff [7].

How many experts should be selected? Experience has shown that the differences among experts can be very important in determining the total uncertainty expressed about a question. Clemen and Winkler [8] examine the impact of dependence among experts using a normal model and conclude that three to five experts are adequate. Hora [9] created synthetic groups from the responses of real experts, and found that three to six or seven experts are sufficient, with little benefit from additional experts beyond that point. When experts are organized in groups, and each group provides a single response, this advice would apply to the number of groups. The optimal number of experts within a group has not been investigated and is likely to be dependent on the complexity of issues being answered.

The Quality of Judgments

Because subjective probabilities are personal and vary from individual to individual and from time to time, there is no “true” probability that one might use as a measure of the accuracy of a single elicited probability. For example, consider the question “what is the probability the next elected president of the United States is a woman?” Individuals may hold different probabilities or degrees of belief about this event occurring. There is, however, no physical, verifiable probability that could be known but remains uncertain. The event will resolve as occurring or not but, will not resolve to a frequency or probability.

It is possible to address the goodness of probabilities, however. There are two properties that are desirable to have in probabilities:

- probabilities should be informative
- probabilities should authentically represent uncertainty.

The first property, being informative, means that probabilities closer to 0.0 or 1.0 should be preferred to those closer to 0.5 as the more extreme probabilities provide greater certainty about the outcome of an event. In a like manner, continuous probability distributions that are narrower or tighter convey more information than those that are diffuse. The second property, the appropriate representation of uncertainty, requires consideration of a set of assessed probabilities. For those events that are given an assessed probability of \( p \), the relative frequency of occurrence of those events should approach \( p \).

To illustrate this idea, consider two weather forecasters who have provided precipitation forecasts as probabilities. The forecasts are given to a precision of one digit. Thus a forecast of 0.2 is taken to mean that there is a 20% chance of precipitation. Forecasts from two such forecasters are shown in Figure 1.

Ideally, each graph would have a 45° line indicating that the assessed probabilities are faithful in that they correctly represent the uncertainty about reality. Weather Forecaster B’s graph shows a nearly perfect relation while the graph for the Forecaster A shows poorer correspondence between the assessed probabilities and relative frequencies with the actual frequency of rain exceeding the forecast probability. The graph is not even monotonic at the upper end.

Graphs showing the relation between assessed probabilities and relative frequencies are called calibration graphs and the quality of the relationship is loosely called calibration which can be good or poor [10]. Calibration graphs can also be constructed for continuous assessed distributions. Following [9], let \( F_i(x) \) be a set of assessed continuous probability
4 Expert Judgment

![Calibration chart](image)

Figure 1 Calibration graph for two forecasters.

distribution values and let \( x_i \) be the corresponding actual values of the variables. If an expert is perfectly calibrated, the cumulative probabilities of the actual values measured on each corresponding distributions function, \( p_i = F_i(x_i) \), will be uniformly distributed on the interval \([0,1]\). We can use the area between the 45° line of perfect calibration and the observed calibration curve as a measure of miscalibration for continuous probability assessments.

Although calibration is an important property for a set of probabilities or probability distributions to possess, it is not sufficient as the probabilities or probability distributions may not be informative. For example, in an area where it rains on 25% of the days, a forecaster who always predicts a 25% chance of rain will be perfectly calibrated but provide no information from day to day about the relative likelihood of rain. But information and calibration are somewhat at odds. Increasing the information by making probabilities closer to zero or one or by making distributions tighter may reduce the level of calibration.

One approach to measuring the goodness of probabilities is through scoring rules. Scoring rules are functions of the assessed probabilities and the true outcome of the event or value of the variable that measure the goodness of the assessed distribution and incorporate both calibration and information into the score. The term **strictly proper scoring function** refers to the property that the expected value of the function is maximized when the probabilities or probability functions to which the function is applied are identical to the probabilities or probability functions that are used to take the expectation. An example will clarify.

A simple strictly proper scoring rule for the assessed probability of an event is the Brier or quadratic rule \([11]\):

\[
S(p) = \begin{cases} 
-(1-p)^2 & \text{if the event occurs} \\
-p^2 & \text{if the complement of the event occurs}
\end{cases}
\]

where \( p \) is the assessed probability. For any probability \( q \), the mathematical expectation

\[
E_q[S(p)] = -q(1-p)^2 - (1-q)p^2
\]

is maximized with respect to \( p \) by setting \( p = q \). Thus, if an expert believes the probability is \( q \), the expert will maximize the perceived expectation by responding with \( q \). In contrast, the scoring rule \( S(p) = -p \) if the event occurs and \( S(p) = -(1-p) \), while intuitively pleasing, does not promote truthfulness. Instead, the expected score is maximized by providing a probability \( p \) of either 0.0 or 1.0 depending on whether \( q \) is less than or larger than 0.5. Winkler [12] provides a discussion of this Brier rule and other strictly proper scoring rules. See also [6, 10].

The concept of a strictly proper scoring rule can be extended to continuous distributions [13]. For example, the counterpart to the quadratic scoring rule for continuous densities is:

\[
S[f(x), w] = 2f(w) - \int_{-\infty}^{\infty} f^2(x) \, dx
\]
Expected scores can sometimes be decomposed into recognizable components. The quadratic rule for continuous densities can be decomposed in the following manner. Suppose that an expert's uncertainty is correctly expressed through the density \( g(x) \), but the expert responds with \( f(x) \) either through inadvertence or intention. The expected score can be written as follows:

\[
E_x [S[f(x), w]] = I(f) - C(f, g)
\]

where \( I(f) = \int_{-\infty}^{\infty} f^2(x) \, dx \) and \( C(f, g) \) is a nonnegative function that increases as \( g(x) \) diverges from \( f(x) \). Thus \( C(f, g) \) is a measure of miscalibration. Further discussion of decomposition can be found in [10, 14, 15]. Haim [16] provides a theorem that shows how a strictly proper scoring rule can be generated from a convex function. See also Savage [17].

**Combining Expert Judgments**

There are two situations that may require probability judgments to be combined. The first is when an expert has provided judgments about the elements of a model that result in a top event probability or distribution of a value. Examples of such decompositions include fault trees, event trees (see Decision Tree), and influence diagrams (see Influence Diagram). With an event or probability tree, the recomposition can be accomplished by simple probability manipulations. In more complicated situations, you may need to employ simulation methods to obtain a top event probability or distribution for a quantity. This is a relatively straightforward process.

Judgments may also be combined across multiple experts. While using multiple experts to address a single question allows for a greater diversity of approaches and often provides a better representation of the inherent uncertainty, doing so creates the problem of having multiple answers when it would be convenient to have a single answer. If you decide to evaluate the risk model separately, using the judgments of each individual expert and you have multiple points in the model where different experts have given their judgments, the number of separate evaluations of the model is the product of the number of experts used at each place in the model and can be very large. Aggregation of the judgments into a single probability or distribution avoids this problem.

There are two classes of aggregation methods, behavioral and mathematical. Behavioral approaches entail negotiation to reach a representative or consensus distribution. Mathematical methods, in contrast, are based on a rule or formula. The approaches are not entirely exclusive, however, as they may be both be used to greater or lesser degree to perform an aggregation.

You may be familiar with the "Delphi" technique, developed at the Rand Corporation in the 1960s by Norman Dalkey [18]. In the Delphi method, the interaction among the experts is tightly controlled. In fact, they do not meet face to face but remain anonymous to one another. This is done to eliminate the influence that one might have because of position or personality. The judgments are exchanged among the experts, along with the reasoning for the judgments. After viewing all the judgments and rationales, the experts are given the opportunity to modify their judgments. The process is repeated—exchanging judgments and revising them—until the judgments become static or have converged to a consensus. Often times, it will be necessary to apply some mathematical rule to complete the aggregation process.

Another behavioral approach is called the **nominal group technique** [19]. This technique, like Delphi, controls the interaction among experts. The experts meet and record their ideas or judgments in written form without discussion. A moderator then asks each expert to provide the idea or judgment and records this on a public media display such as a white board, flip chart, or computer screen. There may be several rounds of ideas/judgments which are then followed by discussion. The judgments/ideas are then ranked individually and anonymously by the experts and the moderator summarizes the rankings. This process can be followed by further discussion and reranking or voting on alternatives.

Kaplan [20] proposes a behavioral method for combining judgments through negotiation with a facilitator. The facilitator and experts meet together to discuss the problem. The facilitator’s role is to bring out information from the experts and interpret
6 Expert Judgment

a “consensus body of evidence” that represents the aggregated wisdom of the group.

A wide range of mathematical methods for combining probability judgments have been proposed. Perhaps the simplest and most widely used is a simple average termed the linear opinion pool [21]. This technique applies equally well to event probabilities and continuous probability densities or distributions. It is important to note that with continuous distributions, it is the probabilities not the values that are averaged. For example, it is tempting, given several medians, to average the medians, but this is not the approach we are referring to. An alternative to the simple average is to provide differential weights to the various experts, ensuring that the weights are non-negative and sum to one. The values of the weights may be assigned by the staff performing the aggregation or they may result from some measure of the experts’ performance. Cooke [6] suggests that evidence of the quality of probability assessments be obtained using training quizzes with questions from the subject matter area addressed by the expert. The experts are given weights based on the product of the $p$ value for the $\chi^2$ test of calibration and the information, as measured by the entropy in the assessments. A cutoff value is used so that poorly calibrated experts are not included in the combination.

The most elegant approach to combining judgments is provided by Morris [22, 23]. In his approach, a decision maker assigns a joint likelihood function to the various responses that a group of experts might provide and a prior distribution for the quantity or event in question. The likelihood function is conditional on the quantity or event of interest. Also see French [24] for a discussion of the axiomatic approaches to combining experts using a Bayesian approach. The decision maker can then develop the posterior distribution for the uncertain quantity or event using Bayes’ theorem (see Bayes’ Theorem and Updating of Belief).

Various mathematical methods for combining probability judgments have different desirable and undesirable properties. Genest and Zidek [25] describe the following property:

**Strong set-wise function property**

A rule for combining distributions has this property if the rule is a function only of the assessed probabilities and maps $[0, 1]^n \rightarrow [0, 1]$. In particular, the combination rule is not a function of the event or quantity in question.

This property, in turn, implies the following two properties:

**Zero set property**

If each assessor, $i = 1, \ldots, n$ provides $P_i(A) = 0$, then the combined result, $P_c(A)$, should also concur with $P_c(A) = 0$.

**Marginalization property**

If a subset of events is considered, the marginal probabilities from the combined distribution will be the same as the combined marginal probabilities.

The strong set property also implies that the combining rule is a linear opinion pool or weighted average of the form

$$P_c(A) = \sum_{i=1}^{n} \alpha_i P_i(A) \quad (4)$$

where the weights, $\alpha_i$, are nonnegative and sum to one.

Another property is termed the independence property is defined by:

$$P_c(A \cap B) = P_c(A)P_c(B) \quad (5)$$

whenever all experts assess $A$ and $B$ as independent events.

But the linear opinion pool does not have this property. Moreover, the linear opinion pool given above cannot be applied successfully to both joint probabilities and the component marginal and conditional probabilities. That is,

$$P_c(A|B)P_c(B) \neq \sum \alpha_i P_i(A|B)P_i(B) \quad (6)$$

except when one of the weights is one and all others are zero, so that one expert is a “dictator”.

The strong set property was used by Dalkey [26] to provide an impossibility theorem for combining rules. Dalkey’s theorem adopts seven assumptions that lead to the conclusion that while conforming to the seven assumptions, “there is no aggregation function for individual probability estimates which itself is a probability function”. Boardley and Wolff [27] argue that one of these assumptions, the strong set property, is unreasonable and should not be used as an assumption.
While the linear rule does not conform to the independence property, its cousin, the geometric or logarithmic rule does. This rule is linear in the log probabilities and is given by

\[ P_c(A) = k \prod_{i=1}^{n} P_i(A)^{\alpha_i} \text{ where } \alpha_i > 0, \sum_{i=1}^{n} \alpha_i = 1 \]  

and \( k \) is a normalizing constant. The geometric rule also has the property of being externally Bayesian.

**Externally Bayesian**

The result of performing Bayes’ theorem on individual assessments and then combining the revised probabilities is the same as combining the probabilities and then applying Bayes’ theorem.

While the geometric rule is externally Bayesian, it is also dictatorial in the sense that if one expert assigns \( P_i(A) = 0 \), the combined result is necessarily \( P_c(A) = 0 \). We note that the linear opinion pool is not externally Bayesian.

It is apparent that all the desirable mathematical properties of combining rules cannot be satisfied by a single rule. The topic of selecting a combining method remains an open topic for investigation.

**Expert Judgment Designs**

In addition to defining issues and selecting and training the expert(s), there are a number of questions that must be addressed concerning the format for a probability elicitation. These include the following:

- the amount of interaction and exchange of information among experts;
- the type and amount of preliminary information that is to be provided to the experts;
- the time and resources that will be allocated to preparation of responses;
- What is the venue – the expert’s place of work, the project’s home, or elsewhere?
- Will there be training, what kind, and how will it be accomplished?
- Are the names of the experts to be associated with their judgments, and will individual judgments be preserved and made available?

The choices result in the creation of a design for elicitation that has been termed a *protocol*. Some protocols are discussed in [6, 28–30]. We briefly outline two different protocols that illustrate the range of options that have been employed in expert elicitation studies.

Morgan and Henrion [28] identify the Stanford Research Institute (SRI) assessment protocol as, historically, the most influential in shaping structured probability elicitation. This protocol is summarized in [31]. It is designed around a single expert (subject) and single analyst engaged in a five-stage process detailed below:

- motivating – rapport with the subject is established and possible motivational biases explored;
- structuring – the structure of the uncertainty is defined
- conditioning – the subject is conditioned to think; fundamentally about his judgment and to avoid cognitive biases
- encoding – this is the actual quantification in probabilistic terms;
- verifying – checking for consistency, the responses obtained in the encoding.

The role of the analyst in the SRI protocol is primarily it to help the expert avoid psychological biases. The encoding of probabilities roughly follows a script. Stael von Holstein and Matheson [4] provide a script showing how an elicitation session might go forward.

The encoding stage for continuous variables is described in some detail in [31]. It begins with assessment of the extreme values of the variable. An interesting sidelight is that after assessing these values, the subject is asked to describe scenarios that might result in values of the variable outside of the interval and to provide a probability of being outside the interval. The process next goes to a set of intermediate values whose cumulative probabilities are assessed with the help of the probability wheel. The probability wheel provides a visual representation of a probability and its complement. Then, an interval technique is used to obtain the median and quartiles. Finally, the judgments are verified by testing for coherence and conformance with the expert’s beliefs.

While the SRI protocol was designed for solitary experts, a protocol developed by Sandia Laboratories for the US Nuclear Regulatory Commission [5, 32] was designed to bring multiple experts together. The Sandia protocol consists of two meetings:
8 Expert Judgment

First meeting agenda
• presentation of the issues and background materials;
• discussion by the experts of the issues and feedback on the questions;
• a training session, including feedback on judgments.

The first meeting is followed by a period of individual study of approximately 1 month.

Second meeting agenda
• discussion by the experts of the methods, models, and data sources used;
• individual elicitation of the experts.

The second meeting is followed by documentation of rationales and opportunity for feedback from the experts. The final individual judgments are then combined using simple averaging to the final probabilities or distribution functions.

There are a number of significant differences between the SRI and Sandia protocols. First, the SRI protocol is designed for isolated experts while the Sandia protocol brings multiple experts together and allows them to exchange information and viewpoints. They are not allowed, however, to view or participate in the individual encoding sessions or comment on one another’s judgments. Second, in the SRI protocol, it is assumed that the expert is fully prepared in that no additional study, data acquisition, or investigation is needed. Moreover, the SRI protocol places the analyst in the role of identifying biases and assisting the expert in counteracting these biases, while the Sandia protocol employs a structured training session to help deal with these issues. In both protocols, the encoding is essentially the same, although the probability wheel is today seldom employed by analysts. Third, the Sandia protocol places emphasis on obtaining and documenting multiple viewpoints which is consistent with the public policy issues addressed in those studies to which it had been applied.

References


Expert Judgment


Related Articles

Subjective Probability

Uncertainty Analysis and Dependence Modeling

Scientific Uncertainty in Social Debates around Risk

STEPHEN C. HORA
10 Expert Judgment

Abstract: Experts are often used to provide uncertainty distributions in risk analyses. They play an important role when insufficient data exist for quantification, or when the available data or models are conflicting. Multiple steps are required in constructing a successful expert judgment process. These steps include selecting and framing the issues, identifying the experts, deciding upon an organization structure, and possibly combining the distributions from multiple experts.

Expert judgments are normally given as probabilities or probability distributions that express the uncertainty about future events or unmeasured quantities. The goodness of probabilistic judgments is measured through calibration and information, which, in turn, can be measured through a scoring rule.

Various behavioral and mathematical methods have been proposed for combining the judgments of experts. There is, however, no single method that has emerged as the best.

Keywords: subjective probabilities; scoring rules; calibration; elicitation; protocol; Delphi

Author Contact Address: University of Hawaii at Hilo, Hilo, HI, USA