# *5 Aleatory Variability and Epistemic Uncertainty*

Aleatory variability and epistemic uncertainty are terms used in seismic hazard analysis that are not commonly used in other fields, but the concepts are well known.

Aleatory variability is the natural randomness in a process. For discrete variables, the randomness is parameterized by the probability of each possible value. For continuous variables, the randomness is parameterized by the probability density function.

Epistemic uncertainty is the scientific uncertainty in the model of the process. It is due to limited data and knowledge. The epistemic uncertainty is characterized by alternative models. For discrete random variables, the epistemic uncertainty is modelled by alternative probability distributions. For continuous random variabiles, the epstemic uncertainty is modelled by alternative probability density functions. In addition, there is epistemic uncertainty in parameters that are not random by have only a single correct (but unknown) value.

The terms randomness and uncertainty have also been used for aleatory variability and epistemic uncertainty, respectively; however, these terms are commonly used in generic ways. As a result, they are often mixed up when used in hazard analysis. The terms "aleatory variability" and "epistemic uncertainty" do not roll off the tongue easily. This unfamiliarity causes people to stop and think about what they are trying to say before using them. The overall goal is to have a clear terminology that will avoid misunderstandings.

## 5.1 Example of the Unknown Die

As a simple example, consider the problem of rolling a die. Assume that you have not seen the die, but you have seen the results of four previous rolls. Those four previous rolls came up 2, 3, 3, and 4. What is the model for this die?

As one approach to developing a model, you may consider that although there are only four observations from this die, you have previous experience with dice. Most dice have six sides that are equally likely. The sparse data set is consistent with a standard die. So you construct a model of the die in which the aleatory is given by a uniform distribution with values between 1 and 6 (model 1 in Table 5-1).

An alternative approach to developing the model could be purely empirical. Given the four observations are all between 2 and 4, you could develop a model that assumes that this is a five-sided loaded die so that it comes up 3 most often, 2 and 4 less often, and 1 and 5 least often. This model is shown as model 2 in Table 51.

Which one of these two models is correct? We don't know until additional data is collected (e.g. more rolls of the die). With time, as more rolls of the die become available, we will be able to distinguish between these two models. These two models represent epistemic uncertainty in the properties of the die.

As additional rolls of the die become available, the aleatory variability does not go to zero. Rather, our estimate of the aleatory variability becomes more accurate. It may increase or decrease from our original estimate. In contrast, the epistemic uncertainty will go to zero as the number of rolls of the die becomes large. If we had a large enough number of rolls, we could develop a very accurate empirical model of the die.

Table 5-1. Example of aleatory variability and epistemic uncertainty for a die with unknown properties.

| | Probability | |
|---|---|---|
| Value | Model 1 | Model 2 |
| 1 | 1/6 | 0.1 |
| 2 | 1/6 | 0.2 |
| 3 | 1/6 | 0.4 |
| 4 | 1/6 | 0.2 |
| 5 | 1/6 | 0.1 |
| 6 | 1/6 | 0.0 |

The alternative models may not be equally credible. In this example, it may be judged to be more likely that the die is a standard six-sided die than some special loaded five-sided

die. The alternative models are assigned weights using logic trees as discussed in detail in section 5.3.

This example demonstrates the situation that is common in developing models for seismic hazard analysis. Often we have a very small amount of data that is from the particular region under study. The two alternatives are to build a model based on the very limited, but region-specific data (e.g. model 2 above) or to use a larger set of data from regions that we consider to be analogous to the region under study (model 1 above).

**5.2 Is it Aleatory Variability or Epistemic Uncertainty?**

The idea of distinguishing between aleatory variability and epistemic uncertainty sounds simple enough and if seismic hazard analyses were about throwing dice it would be easy. In practice, the distinction between aleatory variability and epistemic uncertainty can get confusing.

In distinguishing between aleatory variability and epistemic uncertainty it can be helpful to think how you would describe, in words, the parameter under consideration. If the parameter sometimes has one value and sometimes has another values, then it has aleatory variability. That is, the variability is random. If the parameter always has either one value or another, but we are not sure which it is, then the parameter has epistemic uncertainty.

As an example, consider a fault with a postulated segmentation point. One model may consider that the segmentation point is an impenetrable barrier and an alternative model may consider that the segmentation point does not exist. In this case, the existence of the segmentation point is epistemic uncertainty. There are two alternative segmentation models for the fault. The fault is either unsegmented or it is segmented and always stops the rupture.

Another model of the fault segmentation may consider that the segmentation point exists, but only stops some of the ruptures. In this case, there is aleatory variability in rupture

mode. Sometimes the rupture is stopped at the segmentation point and sometimes the rupture breaks through the segmentation point. In this model, the epistemic uncertainty is in the probability that the rupture will stop at the segmentation point. For example, one model may have a 30% probability that the rupture is stopped at the segmentation point, whereas, another model may have a 10% probability that the rupture is stopped at the segmentation point.

One recurring issue in separating aleatory variability and epistemic uncertainty concerns the limits of what could be learned in the future. There is a school of thought that there is no aleatory variability in the earthquake process. In principle, earthquakes are responding to stresses and strains in the earth. Eventually, given enough time, we will collect enough data to develop detailed models of the earthquake process that give the magnitudes and locations of future earthquakes. Since the earthquake process is in theory knowable, there is only epistemic uncertainty due to our lack of knowledge which will be reduced in time.

A similar issue comes up for ground motion attenuation relations. Ground motion attenuation relations typically only use distance from the site to the source to describe the wave propagation. The detailed 3-D structure of the crust is knowable (or empirical Green's functions could be collected). In principle, with time, a wave propagation model could be determined for each specific source and site. One could argue that the variability of the ground motion attenuation relation that is due to the complexities of the wave propagation should be epistemic uncertainty since it can be determined as additional data become available.

In practise, we have not used this concept of what is potentially knowable long in the future in the estimation of epistemic uncertainty and aleatory variability. Rather, the aleatory variability is determined in the context of the models and is based on the parameterization used in the model. With this approach, a model that only uses distance for the wave propagation will include the variability due to different wave propagation effects as part of the aleatory variability even though it is potentially knowable.

With this approach, the aleatory variability can be reduced as additional fixed parameters are added to the model. For example, consider the case of attenuation relations. In the 1980s, attenuation relations began to include a parameter for the style-of-faulting (e.g. strike-slip or reverse) as part of the model. Since there was a systematic difference in the median ground motion from strike-slip and reverse faults, the inclusion of this factor reduced the standard deviation (aleatory variability) of the attenuation relation. For faults with a single predominate style-of-faulting, then this factor is fixed for future earthquakes and there is a net reduction in the aleatory variability. The penalty for the additional parameter is that in the short term, there is additional epistemic uncertainty as to the value of the style-of-faulting factor term.

What then happens to the models that did not include this additional parameter? If the new parameterization results in a significant reduction in the aleatory variability, then the previous models that did not use this improved parameterization should be down-weighted, until they are finally given zero weight (e.g. they are superseeded).

The addition of parameters to the model that are not fixed for future events does not lead to a reduction of the aleatory variability. For example, consider a new attenuation relation that includes stress-drop as a parameter. The addition of this parameter results in a reduction of the standard deviation determined from a regression analysis; however, since the stress-drop is not fixed for future earthquakes, it must then be randomized for future earthquakes. The result is that there is no systematic reduction in aleatory variability. (There is still a net benfit as the source of the aleatory varibaility is better understood. This is discussed further in sections 6, 7, and 8).

### 5.3 Logic Trees

Epistemic uncertainty is considered by using alternative models and/or parameter values for the source characterization and ground motion attenuation relation. For each combination of alternative models, the hazard is recomputed resulting in a suite of alternative hazard curves. In seismic hazard analyses, it is common to use logic trees to handle the epistemic uncertainty.

A logic tree consists of a series of branches that describe the alternative models and/or parameter values. At each branch, there is a set of branch tips that represent the alternative credible models or parameter. The weights on the branch tips represent the judgment about the credibility of the alternative models. The branch tip weights must sum to unity at each branch point. Only epistemic uncertainty should be on the logic tree. A common error in seismic hazard analyses is to put aleatory variability on some of the branches.

The weights on the branches of logic trees are often called probabilities, but they are better characterized as weights that reflect the current scientific judgments in the relative merit in the alternative models. Calling these weights "probabilities" implies a mathematical basis that does not exist. Epistemic uncertainty is due to limited data (often very limited). In seismic hazard analyses, evaluating the alternative models involves considering alternative simplified physical models, data from analogous regions, and empirical observations. These are subjective. In some cases, uncertainties are developed from statistical evaluations, but that is not usually the case.

Prior to the use of logic tress, the approach was to develop the single best model. In controversial projects, there would be disagreement between the project sponsors, regulators, and interveners as to which model was the best model. Logic trees were used to allow multiple models to be considered with weights that reflected the degree of belief of the scientific community (or at least the seismic hazard analyst) in the alternative models. In this way, all proposed models that were credible could be considered without having to select a single best model.

Using logic trees results in a suite of alternative of the estimates of the hazard each with an associated weight.

### 5.4 Underestimation of Epistemic Uncertainty

In the above discussion, the logic trees are interpreted to represent the scientific uncertainty in the source characterization and ground motion attenuation; however, in practice, logic trees represent the range of available alternative models. In many cases, the range of available models will not cover the epistemic uncertainty. In developing the epistemic uncertainty, the guiding concept should be that less data means larger uncertainty. This seems like a simple concept, yet it is often not followed.

Consider two faults, one that has had many studies and another that has had only a single study. In current practice, the logic tree will consider only available models (sometime this is further restricted to models published in referred journals). So for the well-studied fault, there will be several alternative models available, but for the poorly studied fault there will be only a single model available. By considering only the one available model for the poorly studied fault, 100% weight is given to that model, implying that there is no uncertainty in the model. In contrast, for the well-studied fault, there will be several alternative models available over which the weights will be spread. The result of this approach is that the computed epistemic uncertainty will be larger for the fault with more data.

An additional consequence of the current practice is that as additional research is conducted, additional models will be developed, leading to more branches on the logic tree and larger uncertainty in the computed hazard. Over time this is what has happened. In many cases, our estimates of the epistemic uncertainty have increased, not decreased as additional data have been collected and models developed (ref). This reflects a tendency of scientists to underestimate epistemic uncertainty in regions with little data.

But how can you develop uncertainty estimates with no data? Recall our guiding concept is that less data means larger uncertainty. Regions with more data and more models can

be used as a lower bound of the uncertainty for regions with little or no data. What is needed is a set of generic uncertainties that can be used with simple models to capture the possible range of behaviors of more complex models for regions with little or no data.