PART ONE

STATISTICAL BACKGROUND

CHAPTER ONE

Introduction

UNCERTAINTY IN THE ENGINEERING CONTEXT

1.1 Uncertainty

The engineering function is to design, produce, test, and service structures, devices, materials, and processes that meet a market need, reliably and at a competitive cost. Much of the work involved with that function deals with *mathematical models* of the engineer's designs and the physical phenomena encountered by them. These models efficiently aid the engineer in the design of his artifacts and in the prediction of their behavior.

Consider, for example, the mathematical model for the deflection of the unsupported end of a beam whose other end is rigidly fixed (such a beam is called a "cantilever"). For a beam of length L and cross-sectional moment of inertia I, the deflection D of the free end is

$$D = \frac{WL^3}{3EI},$$

where W is a point load on the free end of the beam and E is a material property called the *modulus of elasticity*. One design problem is to choose the shape and dimensions of the beam cross section so that the deflection is limited to a specified value D.

To use these models, the engineer requires *information* on the constants, parameters, and functional variables that enter the model. The seasoned engineer will be aware of the *uncertainties* that often come with these values. He will therefore know that his designs and predictions will also be associated with uncertainties.

In the above example, the engineer will need to know the magnitude of the load W. Suppose that his design is to support an overhead conveyor track for transporting manufactured parts to an assembly operation. He will notice that W varies from instant to instant. He therefore needs to choose a design value W_D so that his design meets the specified deflection limit D. Similarly, he may realize that the E-value *varies* among material test specimens. He therefore needs to choose a design value E_D so that his design is likely to perform adequately.

How an engineer deals with the uncertainties he faces depends on their relative magnitudes and their likely consequences. If he expects that only small variations of a parameter may occur around a large central value, or if the effect of the expected variation on the performance of the device is small, he may ignore the uncertainty altogether and assume a suitable constant value. If, however, the expected variation is large, or its effect may produce a failure with serious consequences, then the uncertainty is *significant*, and he must deal with it explicitly.

Continuing with the preceding example, the engineer may realize that the E-value for his material is likely to depart from its average value by no more than a few percent. He may therefore declare it to be constant for his purposes and work with its average value or, to be conservative, some low percentile value. However, because the loads W that are likely to occur may vary across an order of magnitude for different parts carried to assembly, he needs to be careful about his choice of design load W_D.

How the engineer accommodates significant uncertainty in his calculations depends on the situation. He may simply take the most detrimental value that he considers possible and multiply it by a "safety factor" to arrive at a deterministic design value. The result is typically a costly overdesign. Furthermore, it is not possible to assess the *risk of failure* for such a design when the extreme cannot be accurately predicted.

In the above example, it may be perfectly acceptable to take the weight of the heaviest part to be transported by the conveyer track and multiply it by the factor of, say, two to obtain the design load W_D. The increased cost for a few such cantilevers is unlikely to be significant. However, if the design is to be mass produced, the increased cost *will* be significant, so that an arbitrary safety factor may not be acceptable. Similarly, if the maximum load W for a range of design applications cannot be predicted, and only sample values of possible loads are available, the likelihood of adequate performance of the design cannot be assessed under this deterministic approach.

The modern climate of competitiveness demands the efficient use of materials, as well as more reliable designs. The former requirement pushes the design closer to the possibility of failure, whereas the latter demand obviously pushes in the opposite direction. To deal with this dilemma in a professional manner, the engineer must treat the significant uncertainties of his information base. The *rational* approach to dealing with such uncertainties is to construct a *mathematical model* that describes the uncertainty aspect of engineering information, similar to representing a physical phenomenon. The *measure* of this uncertainty aspect of engineering quantities is *probability*. This text deals specifically with some useful ways to model the uncertainty and variation associated with the quantitative information encountered in a wide variety of engineering contexts.

1.2 Sources of Uncertainty

The question might be asked at this point: "If engineering information is often uncertain, where does that uncertainty come from?" The *sources* of uncertainty can be broadly classified as follows.

1. data uncertainty,
2. statistical uncertainty,
3. event uncertainty, and
4. model uncertainty.

Data Uncertainty

The majority of the quantities the engineer measures or observes feature *inherent variability*. That is, the measured value is caused, or influenced, by many chance factors whose effects aggregate to produce a measured value. No matter how carefully one measures such a quantity, variability among measured values is an inherent reality, so that the actual value of a future measurement is uncertain.

Examples of inherently variable quantities are: the yield of a chemical batch process, the specific strength of an engineering material, the gust load on an aircraft wing, the time-to-failure of equipment, the cost of an engineering project, the duration of an assembly task, the throughput time of a production order, the propagation rate of a combustion flame front, and the number of stress reversals to the fatigue failure of a metal specimen.

Inherent variability is a source of uncertainty the engineer encounters commonly. However, even when the engineer measures a quantity that is inherently constant, such as the distance between two survey stations, he will find that his observations vary at the limit of precision of his measuring instruments. That is, the process of measurement itself often introduces uncertainty, regardless of whether the measured quantity is inherently constant or variable. This *measurement uncertainty* may show up as instrument *bias* where the true value differs from its measured value by some consistent amount. *Calibration* of the instrument by a standard of sufficient accuracy reduces this uncertainty. Measurement uncertainty also appears in the form of *random* differences between true and measured values, in much the same way that inherent variability occurs. Using instruments of sufficient *precision* and carefully designing the instrumentation system render this uncertainty insignificant compared to the precision required in the measurement.

Statistical Uncertainty

An important source of uncertainty is the limited amount of information that is typically available on a measurable quantity. That is, the quantity of interest may be measurable an unrestricted number of times, but time and budget constraints permit only few observations. Clearly, the more observations are on hand, the more information on that quantity is available. Thus, *limited* information implies uncertainty about the true nature of the quantity.

For example, if only five prototypes of a newly developed device are available for performance testing, the project engineer could only form a rather approximate impression of the design's performance. A characteristic such as average performance would then be highly uncertain. Other important characteristics, such as 5-percentile performance, could not even be expressed sensibly. If, however, one had fifty prototypes to test, one could draw conclusions on design performance with more assurance that they are close to true values.

An assessment of statistical uncertainty is made in the form of *standard errors*, or related measures, attached to the predictions of the quantity in question. Error statements, and the like, are important means of communicating the presence and magnitude of uncertainty to the decision maker who can then evaluate the *risk* associated with his decision. In this text the methods for constructing these uncertainty measures are presented and illustrated.

Event Uncertainty

In specifying his design, the engineer will have to guard against the effects of unfavorable events that may occur during the design's mission. That is, he is concerned with the *occurrence* of events. The events of design significance are usually of the kind that happen only rarely, so that typically there is little information available on their likely occurrence. The resulting design decision is often highly uncertain.

For example, the engineer would consider the possibility of a major meteorite impact in the design of a spacecraft structure. He would consider the chance of a high wind load occurring during the construction phase of a suspension bridge. In the design of a communications tower he would consider the possibility of a high load induced by a major earthquake.

Model Uncertainty

The mathematical models the engineer uses in his work typically represent only one, or a few, of the important features of the physical phenomenon in question. That is, a model represents a restricted version of reality. Furthermore, the model's description of a real problem is often an idealization. The model therefore deviates from reality. When the model's "lack of fit" to reality significantly affects the conclusions drawn from it, these conclusions are in error. To reduce this type of uncertainty, one needs to construct more realistic models.

For example, material *failure* mechanisms involve complex interactions of many contributing causes. Mathematical models of failure only consider one or two of the important factors and relate these to failure events by simple algebraic relations. Predictions of material behavior based on these models often differ widely from what is observed. In contrast, although many strain models are

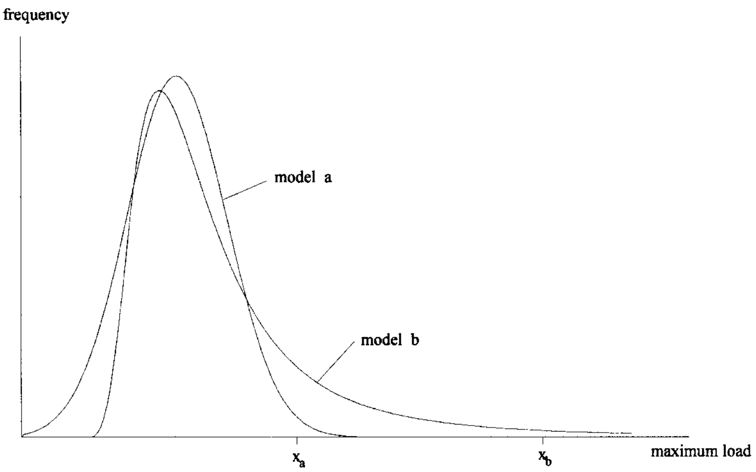

Figure 1.1. Two uncertainty models of a load variable.

linear idealizations of nonlinear reality, their predictions are quite accurate, at least for small strains.

The problem of model-fit, when describing physical phenomena, also appears when a *probability model* is chosen to describe the uncertainty aspect of engineering information. Again, when the model does not accord with reality, the conclusions from the model are in error.

For example, suppose that an engineer has on hand a good set of observations on the maximum of a load to which his equipment design will be exposed during its mission. Suppose further that he needs to know the 99-percentile maximum load as an input to his design process. Figure 1.1 shows two probability models representing the maximum load variable. The two models have about the same average and spread, in line with the data, and both are acceptable as fair representations of the bulk of the data. However, the estimate x of the 99-percentile load differs by a factor of about *two* between the models. Thus, the question of which model provides a better description of this load phenomenon clearly needs further and careful attention.

This text alerts the practicing engineer to a variety of useful models available to him, and it describes the practical situations in which each model is appropriate.

1.3 Population versus Sample

It is useful to explore the notion of *statistical uncertainty* further to make the important distinction between a *population* of individuals and a *sample* of individuals from that population. When all individuals of a population are known (i.e., when all possible measurements on a variable are on hand), one has *complete information* on that population (i.e., the true nature of the measured phenomenon is known). In practice, one cannot usually know all individuals in a population. Instead, only a small selection of individuals (a sample) is known. The sample thus represents *incomplete information* on the population: It is all the quantitative information one has. However, one naturally wishes to form conclusions about the population, given the incomplete information of the sample. Conclusions, then, pertain to populations (i.e., a measurement phenomenon in general), whereas the available information base only comprises the knowledge residing in the sample (i.e., the limited set of actual observations on hand). Thus, sample information is fully known, but conclusions about the population are necessarily shrouded in uncertainty.

For example, the engineer may have a sample of twenty-five material test specimens, loaded to failure. There is nothing uncertain about the sample itself, barring experimental error. However, with respect to the strength property of this type of material, the sample information is incomplete. Nevertheless, given this incomplete information, the engineer wishes to draw conclusions on the strength of this material in general.

1.4 Statistics versus Probability

The specific measurements in a sample are the data that comprise the engineer's quantitative information base. Given that information, he needs to construct a probability model that describes the variability of the population as accurately as possible. The process of constructing such a model on the basis of data is termed "statistical inference." The main inferential procedures are point estimation, confidence interval estimation, and hypothesis testing (see Chapter 3). These procedures are *inductive*, as they proceed from the specific (sample) to the general (model), and belong to the subject of *statistics*. They answer questions about the population.

Continuing with the preceding example, the engineer may need to establish, on the basis of the twenty-five test results, whether this material is substantially stronger than a specified value. He would do this by means of a hypothesis test.

Once a population model is inferred (or assumed), it forms the basis for answering questions about a *sample*. The procedures used are essentially the rules of *probability*. These rules are *deductive* as they proceed from model to experimental data.

In the above example, suppose the engineer has inferred a probability model of material strength from the data. Given that model, he could deduce the probability that a further test specimen breaks below a specified strength value.

Figure 1.2 shows the distinctions involved. This text focuses on the *statistics* of inferring, from measurement data, the specifics of a distribution, chosen from a number of models that are prominent in engineering and the sciences.

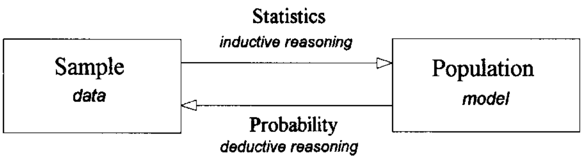

Figure 1.2. The relation between sample and population.

DATA

1.5 Sample Measures

An *experiment* in which a variable of interest X is measured or observed n times produces the data x_i that the engineer works with. Such an experiment may be set up in a laboratory specifically to produce the required data, or it may take place in the field and consist of gathering data as they happen.

For example, a fatigue experiment in a materials testing lab measures the number of stress reversals to failure for several specimens. These measured stress reversals are the *data*. Keeping records of operating times between failures of transmissions for a fleet of trucks produces information on the *life* of these transmissions. These times-to-failure are the data.

The collection of *raw* data in a *sample*,[1] denoted by $\{x_i\}_n$, is by itself not particularly informative. To bring out salient features of the sample, the data are processed to generate *descriptive statistics*, usually in the following sequence:

1. The data are *ordered* according to increasing magnitude, resulting in a rearranged set $\{x_{(i)}\}_n$. Thus, $x_{(1)}$ is the smallest observation in the sample, $x_{(2)}$ is the next-larger observation, and so on. These rearranged values are called "order statistics." The *range* of the data, given by $[x_{(n)} - x_{(1)}]$, tells how widely dispersed the data are.
2. Some useful *sample measures* are calculated, such as the
   - sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
   - sample mode: $x_m =$ most frequent measurement,
   - sample median: $\tilde{x} =$ middle-most ordered measurement,
   - sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,
   - sample standard deviation: $s = \sqrt{\text{sample variance}}$,
   - sample coefficient of variation: $scv = s/\bar{x}$.

   These measures summarize the sample information by describing (a) where the bulk of the data are located on the measurement axis (i.e., $\bar{x}$, x_m, $\tilde{x}$) and (b) how dispersed the sample is (i.e., s (in units of X); scv (dimensionless).

[1] The statistical procedures of this text assume that the sample is "random." This means that the data are drawn independently from a common population; see also Section 2.2.

3. If the sample size n is sufficiently large (at least 25), it is instructive to *group* the data into (usually equal) measurement intervals, so that a sample *frequency distribution* (called a *histogram*) can be displayed in a table or on a graph. That is, if the data range is split into k intervals Δx_j, the frequencies

   $$q_j = \text{the number of observations } x \text{ in } \Delta x_j$$

   can be obtained. The *relative* frequencies

   $$f_j = \frac{q_j}{n}$$

   describe in some detail the variability among the data $\{x_i\}_n$. The *cumulative relative* frequencies

   $$F_j = \sum_{i=1}^{j} f_i$$

   indicate the *distribution* of the likelihood of the values x_i over the data range. The graphical display of f_j provides an informative picture of the variation among the data x_i. See Example 1.1 for a Mathcad document where a sample of 29 measurements is processed as described above.

MODELS

1.6 Distribution Function

When one cannot know in advance the values of repeated measurements on a quantity of interest, it is practical to describe that quantity as a *random variable*, denoted by a capital letter such as X. This random variable refers to the *population*. The collection of all possible values in the population is called the *sample space* S. The *realizations* of X, that is, measurements on X, are denoted by lowercase letters such as x. A set of these realizations is the *sample*. Thus, a sample is a subset of values in the larger sample space S.

The *uncertainty* aspect of the random variable X is modeled by a statistical distribution $F(x;\theta)$. Here F is a mathematical function of the values x that the variable X can take in its sample space S. The model F is indexed by parameters θ. Thus, $F(x;\theta)$ comprises a *family* of distributions, indexed by the values θ can take in its *parameter space* Ω. See Chapters 4 and 9 for a little more detail on the nature of the distribution function F.

For example, suppose an engineer is interested in a random variable X that represents the yield of a chemical process under specified process conditions. Observations x on process yield are realizations, and a set of these constitutes a sample.