12. *Entropy: Modeling Uncertainty*

Information is the resolution of uncertainty.

—Claude Shannon

In this chapter, we introduce *entropy*, a formal measure of uncertainty. With it, we can show an equivalence between uncertainty, information content, and surprise. Low entropy corresponds to low uncertainty and little information being revealed. When an outcome occurs in a low-entropy system, such as the sun rising in the east, we experience little surprise. In high-entropy systems, say the drawing of numbers in a lottery, the outcomes are uncertain and when realized, they reveal information. We experience surprise.

Using entropy, we can compare disparate phenomena. We can say whether election outcomes in New Zealand are more uncertain than outcomes of United Nations votes on censure. We can compare the uncertainty of stock prices to the uncertainty of outcomes of sporting events. We can also use entropy to distinguish between the four classes of outcomes: equilibrium, periodicity, complexity, and randomness. We can distinguish complex patterns that appear random from true randomness and discern whether what appears to be a pattern is, in fact, random.

We can also use entropy to characterize distributions. In the absence of a controlling or regulating force, some populations may drift toward maximal entropy. Given constraints, such as a fixed mean or variance, we can solve for maximum entropy distributions. Maximal entropy results can also guide our modeling choices by justifying some distributions over others.

The chapter has five parts. In the first part, we provide intuition for and define information entropy. In the second part, we describe Shannon's axiomatic foundations for a general class of entropy measures. In the third part, we discuss how to use entropy to distinguish equilibrium, order, randomness, and complexity. In the fourth part, we investigate systems that produce maximal entropy given constraints. We conclude by discussing why, sometimes, we prefer complexity to equilibrium.

Information Entropy

Entropy measures the uncertainty associated with a probability distribution over outcomes. It therefore also measures surprise. Entropy differs from variance, which measures the dispersion of a set or distribution of numerical values. Uncertainty correlates with dispersion, but the two differ. Distributions with high uncertainty have nontrivial probabilities over many outcomes. Those outcomes need not have numerical values. Distributions with high dispersion take on extreme numerical values.

The distinction can be seen in stark relief by comparing a distribution that has maximal entropy with one that has maximal variance. Given outcomes that take values 1 through 8, the distribution that maximizes entropy places equal weight on each outcome.[1] The distribution that maximizes variance takes value 1 with probability image and value 8 with probability image , as shown in figure 12.1.

image

Figure 12.1: Maximal Entropy and Maximal Variance

Entropy is defined over probability distributions. It can therefore be applied to distributions over nonnumerical data such as the species of birds in a forest or the market shares of flavors of jam. The formal expression for entropy is written as minus the sum of products of probabilities and their logarithms. That sounds complicated, but it will become intuitive.

We begin with the special case of *information entropy*, which measures uncertainty in terms of number of random flips of a fair coin. Suppose that every family has exactly two children and that boys and girls are equally likely. The sexes of a family's children (listed by birth order) are equivalent to two coin flips. The distribution over outcomes therefore has an information entropy of 2 because it corresponds to 2 random events. The information content also equals 2 because we could learn the outcomes by asking 2 yes-or-no questions.

Similarly, the sexes of the children in families of size three are equivalent to 3 coin flips. To learn about a family's children, we would need to ask 3 questions. The same logic applies for any number of children. In the general case, to learn the sexes of N children, we would need to ask N questions.

Notice that those N questions distinguish among 2^N possible birth orders. That mathematical relationship is the key to understanding the entropy measure: N binary random events produce 2^N possible outcome sequences, and, equivalently, we could learn the outcome sequence by asking N questions. For this reason, information entropy assigns an uncertainty level (and an information content) of N to an equal distribution over 2^N outcomes.

To capture that relationship in formal mathematics, we first note that each of the outcome sequences has a probability of image. To convert this to N requires the rather complicated expression image.[2] We can generalize this construction to arbitrary probabilities. If an outcome sequence arises with probability p, then we assign an uncertainty $\log_2(p)$ which approximates the number of yes-or-no questions required to identify the sequence To compute the information entropy of a distribution, we average the expected number of questions across all outcomes, or, as in the example, sequences of outcomes.

Information Entropy

Given a probability distribution $(p_1, p_2, \ldots p_N)$, the **information entropy**, H_2, equals:

image

Note: the subscript 2 denotes the use of the base 2 logarithm.

At first, the mathematical representation complicates more than it clarifies. Working through an example makes the formula more intuitive. Imagine that families who first have a girl stop having children, and that families who first have a boy have two more children. Half of all families will have a single girl. The half will be split evenly among four outcomes: three boys, two boys followed by a girl, a boy followed by two girls, and a boy followed by a girl followed by another boy. Each of those four outcomes occurs with probability image.

Information entropy equals the expected number of questions we must ask to learn the family's children. We would first ask if the first child is a girl. With probability image the answer is yes, and we need not ask more questions. Thus, half of the time, we ask one question. We can write this as image. If the answer is no, we must ask two more questions for a total of three questions. Each of those four cases occurs with probability image, so each contributes image × 3 to information entropy. We write each as image. Information entropy equals 2, the sum of the five terms.[3] Notation and logarithms aside, the intuition should be clear: information entropy corresponds to the expected number of yes-or-no questions. If we have to ask a lot of questions, the distribution is uncertain. Knowing the outcome reveals information.

Axiomatic Foundations of Entropy

Axiomatic Foundations: Entropy

image

The above class of **entropy measures** uniquely satisfies the following four axioms:

Symmetric, continuous function: H(σ(image)) = H(image) for any σ that permutes the probabilities.

Maximization: H(image) is maximized at p_i = image for all N.

Zero Property: $H(1, 0, 0,\ldots, 0) = 0$.

Decomposability: If image

image

where image and image

To arrive at a general expression for entropy, we take an axiomatic approach. Claude Shannon imposed four conditions on his measure. The first three are easy to understand. It needed to be continuous and symmetric, maximized when outcomes occur with equal probability, and equal zero for certain outcomes. The fourth condition (decomposability) requires that the entropy of a probability distribution defined over *n* categories each with *m* subcategories equals the entropy of the distribution over the categories plus the sum of the entropies of each of the subcategories. This is a natural assumption for products of distributions. For example, in the case where outcomes are the product of two independent events, the assumption implies that the information content of the joint event equals the sum of the information contents of each event separately. Shannon then proved that a general class of *entropy measures* uniquely satisfies those axioms.

As was the case for the axioms that characterize Shapley values, the contribution of these axioms resides less in their existence than in their reasonableness. A clever mathematician can always construct axioms that uniquely define a function. The first two axioms are difficult to question. We might quibble with the arbitrariness of setting the uncertainty of a known distribution at zero, but it is an appropriate benchmark. Another possibility would be to assign 1 as the uncertainty of a known distribution.[4] The decomposability axiom, though complicated to explain, is also difficult to challenge. The uncertainty of two combined random events should equal the sum of the uncertainties of each event. Overall, the axioms are more than defensible. They are, in fact, hard to dispute.

Using Entropy to Distinguish Classes of Outcomes

We now show how the entropy measure can help us to categorize empirical data and model output within Wolfram’s four classes: *equilibrium, cyclic (periodic), random,* and *complex*.[5] In Wolfram’s classification, a pencil resting on a desk is in equilibrium. The planets orbiting the sun are in a cycle. A sequence of coin flips is random, so are (approximately) stock prices on the New York Stock Exchange, as we shall learn in the next chapter. Finally, the neuronal firings in a person’s brain are complex; they do not fire randomly, nor do they fire in a fixed pattern. Figure 12.2 represents these four categories graphically.

Equilibrium outcomes have no uncertainty, and therefore, have an entropy equal to zero. Cyclic (or periodic) processes have low entropy that does not change with time, and perfectly random processes have maximal entropy. Complexity has intermediate entropy—it lies between ordered and random. While entropy gives us a definitive answer in the two extreme cases, equilibrium and random, it does not for cyclic and complex outcomes. We will have to use other measures to distinguish those cases.

image

Figure 12.2: Wolfram’s Four Classes

To classify a time series of data, we calculate the information entropy across subsequences of different lengths. Suppose that a man keeps track of the type of hat he wears each day—either a beret (B) or a fedora (F). His choices over a year create a binary time series of 365 events. We can first calculate the entropy of sequences of length 1, that is, we calculate the entropy over the probability of wearing each type of hat. If we find that he is equally likely to wear each type of hat, the entropy over sequences of length 1 equals 1. We can therefore rule out equilibrium, as he changes his choices, but any of the other three categories are possible.

To determine the category, we next compute the entropy of sequences of length 2 through 6. If all have maximal entropy, then we can rule out a simple cycle. Suppose that as we consider longer sequences the entropy increases slowly until it reaches a maximum of 8. In other words, no matter how long the subsequence, the entropy never exceeds 8. An entropy of 8 is equivalent to an equal distribution across 256 outcomes. That cannot be a simple cycle. It is more representative of a complex sequence containing structure and patterns. We cannot say for sure that the time series is complex. It might be that the person is trying to be random, yet fails.

Maximal Entropy and Distributional Assumptions

Many of the situations that we model include uncertainty, and, as modelers, we must make assumptions about those distributions. As a rule, we want to avoid making ad hoc assumptions. It may be that we have some understanding of the process that produces the distribution. If so, we can often derive the statistical structure produced by that process using our logic-structure-function approach.

For example, suppose that we want to make an assumption about the distribution of the total value of the items up for auction at an estate sale. The total value equals the sum of the values of the individual items. We can therefore invoke the central limit theorem and assume a normal distribution. We might also assume a normal distribution for the possible values of a house, as the house's value depends on its attributes: the number of bedrooms, bathrooms, and the size of the lot.

A normal distribution may not make sense for the possible values for a piece of art or a rare manuscript. In those cases, we may have little understanding of the process that determines value. One approach is to assume a distribution with maximal uncertainty, that is, the maximal entropy distribution.

The shape of the maximal entropy distribution depends on the constraints. As we have already seen, if we assume a minimal and maximal value, the *uniform distribution* maximizes entropy. Many social science models in textbooks and journals assume uniform distributions. We might question that assumption on the grounds that few distributions in the real world are uniform. However, a *principle of indifference*—if we know nothing other than the range or set of possibilities—can justify the uniform distribution.

In some cases, we may know the mean of the distribution and also know that all values must be positive. Given those constraints, the maximal entropy distribution must have a long tail, and as we spread the distribution across more values, we must balance high values with many low-value outcomes. It can be shown that the entropy-maximizing distribution will be an *exponential distribution*. Thus, if we are writing a model that assumes distribution of website hits or market shares, in the absence of data an exponential distribution is a natural assumption.

Finally, if we fix the mean and the variance (and allow negative values), then the maximal entropy distribution is the normal distribution. The logic here is similar to the previous case. To create more uncertainty, we create extreme values. Here we can balance positive and negative values and not change the mean. However, doing so increases the variance, so we must add more values near the mean, resulting in a bell curve.

We can interpret these maximal entropy distributions within the logic-structure-function framework. If we thought that in a given social, biological, or physical context a micro-level process was maximizing entropy, then we should expect one of these distributions. Alternatively, we might assume a micro-level process and be able to show that entropy increases. If so, one of these distributions would emerge.

Maximal Entropy Distributions

Uniform distribution: Maximizes entropy given a range, $[a, b]$.

Exponential distribution: Maximizes entropy given a mean, μ.

Normal distribution: Maximizes entropy given a mean, μ, and a variance, σ^2.

We can also interpret these results as exploratory. We may encounter data that is exponentially or normally distributed. Though we are not obliged to ask if some underlying behavior is increasing entropy subject to a constraint, we might gain a novel insight by doing so. Previously, we explained the normal distribution of heights, weights, and lengths of species by an appeal to the central limit theorem. Here we present a different, model-based explanation. If mutation maximizes entropy (to best explore niches), and if average size and total dispersion are fixed, then the distribution of sizes will be normal. The point is not that the maximal entropy approach offers a better explanation, but that maximizing entropy given constraints results in a normal distribution. So, when we see a normal distribution, it could be the result of entropy maximization.

Positive and Normative Implications of Entropy

We have seen how entropy measures uncertainty, information, and surprise, how it differs from variance, which measures dispersion, and how it can help us classify and compare classes of outcomes. Later, in Chapters 13 and 14, when we study random walks and path dependence, we use entropy to identify randomness and to measure the extent of path dependence. We can put the entropy measure to use in any number of real-world applications. We can measure whether an intervention in financial markets increases or decreases uncertainty. We can test whether or not outcomes in elections, sporting events, or games of chance are random.

In each of these applications, entropy functions as a positive measure. It tells us what the world is, not what it should be. Entropy in a system is not intrinsically bad or good. How much entropy we desire depends on the situation. In constructing a tax code, we might want an equilibrium pattern of behaviors. We would not want randomness. In designing a city, we may seek complexity. Equilibrium or even cycles would be dull. We would prefer a city to be teeming with life, to offer opportunities for fortuitous meetings and interactions. More entropy would be better, but only to a point. We would not want randomness. Randomness would make planning difficult and possibly overwhelm our cognitive abilities. Ideally, the world produces some complexity and we live in interesting times.

The architect Christopher Alexander shows how geometric properties such as strong centers, thick boundaries, and non-separateness can produce complex, living buildings, neighborhoods, and cities.[6] Alexander argues for complexity in cities and in living space. Central bankers may be less fond of complexity. They may prefer predictable equilibrium outcomes and stable growth paths. A central takeaway from this chapter is that we often care whether a system goes to equilibrium, produces a pattern or randomness, or whether it results in complex, novel sequences of patterns. By using models, we can perhaps see which will arise and, in some cases, design systems that produced the class of outcome we desire, whether that be complexity or equilibrium.