EXCERPTED FROM

STEPHEN WOLFRAM

A NEW KIND OF SCIENCE

SECTION 10.9

Statistical Analysis

tell, the only kinds of correlations that are ultimately important to our auditory system are those that lead to some form of repetition.

So in the end, any features of the behavior of a system that go beyond pure repetition will tend to seem to our ears essentially random.

Statistical Analysis

When it comes to studying large volumes of data the method almost exclusively used in present-day science is statistical analysis. So what kinds of processes does such analysis involve? What is typically done in practice is to compute from raw data various fairly simple quantities whose values can then be used to assess models which could provide summaries of the data.

Most kinds of statistical analysis are fundamentally based on the assumption that such models must be probabilistic, in the sense that they give only probabilities for behavior, and do not specifically say what the behavior will be. In different situations the reasons for using such probabilistic models have been somewhat different, but before the discoveries in this book one of the key points was that it seemed inconceivable that there could be deterministic models that would reproduce the kinds of complexity and apparent randomness that were so often seen in practice.

If one has a deterministic model then it is at least in principle quite straightforward to find out whether the model is correct: for all one has to do is to compare whatever specific behavior the model predicts with behavior that one observes. But if one has a probabilistic model then it is a much more difficult matter to assess its validity—and indeed much of the technical development of the field of statistics, as well as many of its well-publicized problems, can be traced to this issue.

As one simple example, consider a model in which all possible sequences of black and white squares are supposed to occur with equal probability. By effectively enumerating all such sequences, it is easy to see that such a model predicts that in any particular sequence the fraction of black squares is most likely to be $1/2$.

But what if a sequence one actually observes has 9 black squares out of 10? Even though this is not the most likely thing to see, one certainly cannot conclude from seeing it that the model is wrong. For the model does not say that such sequences are impossible—it merely says that they should occur only about 1% of the time.

And indeed there is no meaningful way without more information to deduce any kind of absolute probability for the model to be correct. So in practice what almost universally ends up being done is to consider not just an individual model, but rather a whole class of models, and then to try to identify which model from this class is the best one—as measured, say, by the criterion that its likelihood of generating the observed data is as large as possible.

For sequences of black and white squares a simple class of models to consider are those in which each square is taken to be black with some fixed independent probability p. Given a set of raw data the procedure for finding which model in this class is best—according, say, to the criterion of maximum likelihood—is extremely straightforward: all one does is to compute what fraction of squares in the data are black, and this value then immediately gives the value of p for the best model.

So what about more complicated models? Instead of taking each square to have a color that is chosen completely independently, one can for example take blocks of squares of some given length to have their colors chosen together. And in this case the best model is again straightforward to find: it simply takes the probabilities for different blocks to be equal to the frequencies with which these blocks occur in the data.

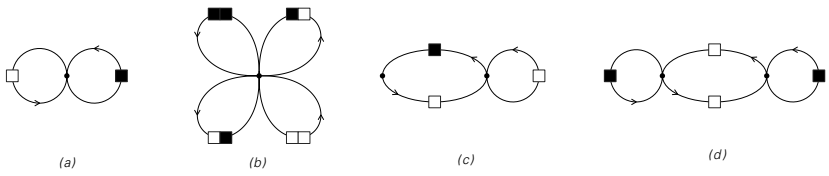
If one does not decide in advance how long the blocks are going to be, however, then things can become more complicated. For in such a case one can always just make up an extreme model in which only one very long block is allowed, with this block being precisely the sequence that is observed in the data.

Needless to say, such a model would for most purposes not be considered particularly useful—and certainly it does not succeed in providing any kind of short summary of the data. But to exclude models like this in a systematic way requires going beyond criteria such as

maximum likelihood, and somehow explicitly taking into account the complexity of the model itself.

For specific types of models it is possible to come up with various criteria based for example on the number of separate numerical parameters that the models contain. But in general the problem of working out what model is most appropriate for any given set of data is an extremely difficult one. Indeed, as discussed at the beginning of Chapter 8, it is in some sense the core issue in any kind of empirical approach to science.

But traditional statistical analysis is usually far from having to confront such issues. For typically it restricts itself to very specific classes of models—and usually ones which even by the standards of this book are extremely simple. For sequences of black and white squares, for example, models that work as above by just assigning probabilities to fixed blocks of squares are by far the most common. An alternative, typically viewed as quite advanced, is to assign probabilities to sequences by looking at the paths that correspond to these sequences in networks of the kind shown below.

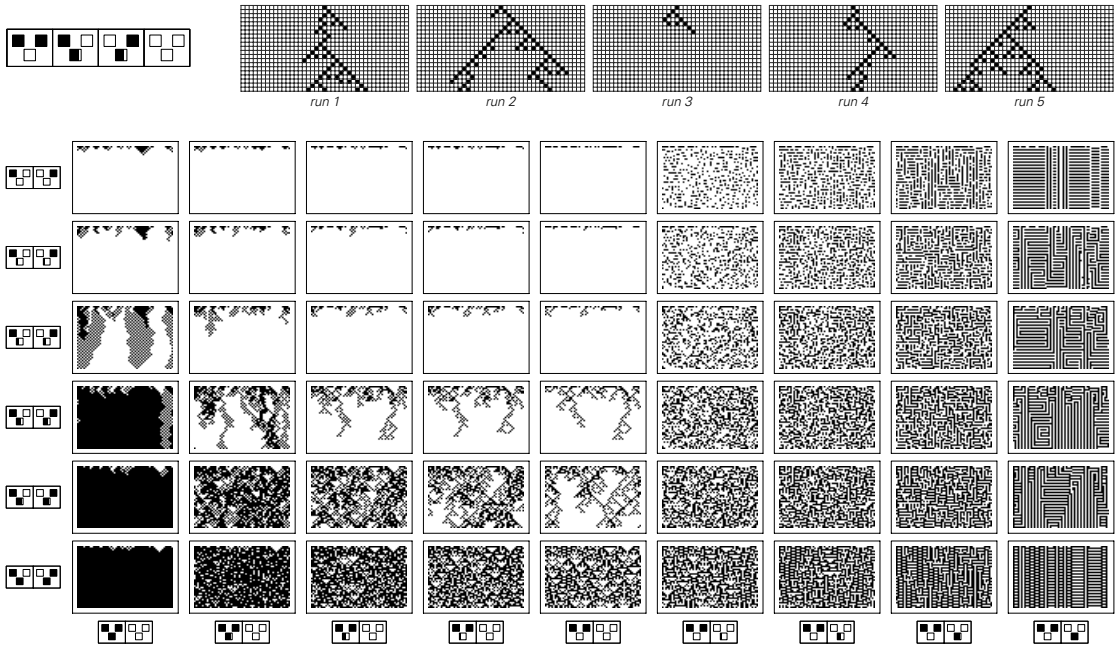


Networks defining probabilistic models. Each connection in each network has a certain probability associated with it, and the model takes sequences of black and white squares to be generated by tracing paths through the networks according to these probabilities. Cases (a) and (b) are so-called Markov models that in effect involve no memory and are equivalent to models discussed above. Cases (c) and (d) correspond to so-called hidden Markov models, with some short-term memory.

Networks (a) and (b) represent cases already discussed above. Network (a) specifies that the colors of successive squares should be chosen independently, while network (b) specifies that this should be done for successive pairs of squares. Network (c), however, specifies that different probabilities should be used depending on whether the path has reached the left or the right node in the network. But at least

so long as the structure of the network is kept the same, it is fairly easy even in this case to deduce from a given set of data what probabilities in the network provide the best model for the data—for essentially all one need do is to follow the path corresponding to the data, and see with what frequency each connection from each node ends up being used.

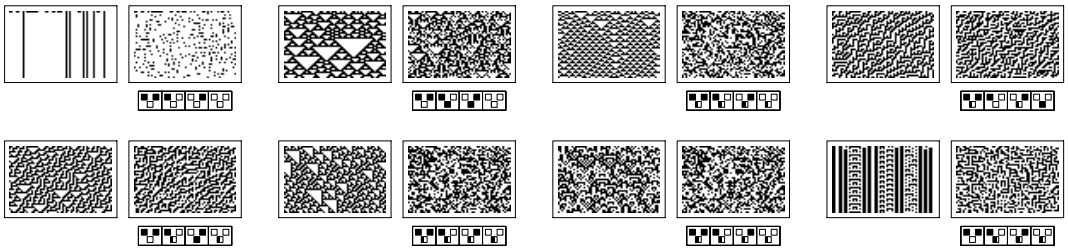
So what about two-dimensional data? From the discussion in Chapter 5 it follows that no straightforward analogs of the types of probabilistic models described above can be constructed in such a case. But as an alternative it turns out that one can use probabilistic versions of one-dimensional cellular automata, as in the pictures below.



Examples of probabilistic cellular automata, in which the rule specifies the probabilities for each color of cell to be generated given what the colors of its two neighbors were on the previous step. Because the rule is probabilistic a different detailed pattern of evolution will in general be obtained each time the cellular automaton is run—as in the top row of pictures above. Despite this, however, any particular probabilistic cellular automaton will typically exhibit some characteristic overall pattern of behavior, as illustrated in the array of pictures above. Note that it is fairly common for phase transitions to occur, in which continuous changes in underlying probabilities lead to discrete changes in typical behavior. Probabilistic cellular automata can be viewed as generalizations of so-called directed percolation models.

The rules for such cellular automata work by assigning to each possible neighborhood of cells a certain probability to generate a cell of each color. And for any particular form of neighborhood, it is once again quite straightforward to find the best model for any given set of data. For essentially all one need do is to work out with what frequency each color of cell appears below each possible neighborhood in the data.

But how good are the results one then gets? If one looks at quantities such as the overall density of black cells that were in effect used in finding the model in the first place then inevitably the results one gets seem quite good. But as soon as one looks at explicit pictures like the ones below, one immediately sees dramatic differences between the original data and what one gets from the model.



A comparison between data generated by ordinary cellular automata and the probabilistic cellular automata that are considered the best fit to it. While properties such as the density of black cells are typically set up to agree between the data and the model, the pictures make it clear that more detailed features do not.

In most cases, the typical behavior produced by the model looks considerably more random than the data. And indeed at some level this is hardly surprising: for by using a probabilistic model one is in a sense starting from an assumption of randomness.

The model can introduce certain regularities, but these almost never seem sufficient to force anything other than rather simple features of data to be correctly reproduced.

Needless to say, just as for most other forms of perception and analysis, it is typically not the goal of statistical analysis to find precise and complete representations of data. Rather, the purpose is usually just

to extract certain features that are relevant for drawing specific conclusions about the data.

And a fundamental example is to try to determine whether a given sequence can be considered perfectly random—or whether instead it contains obvious regularities of some kind.

From the point of view of statistical analysis, a sequence is perfectly random if it is somehow consistent with a model in which all possible sequences occur with equal probability.

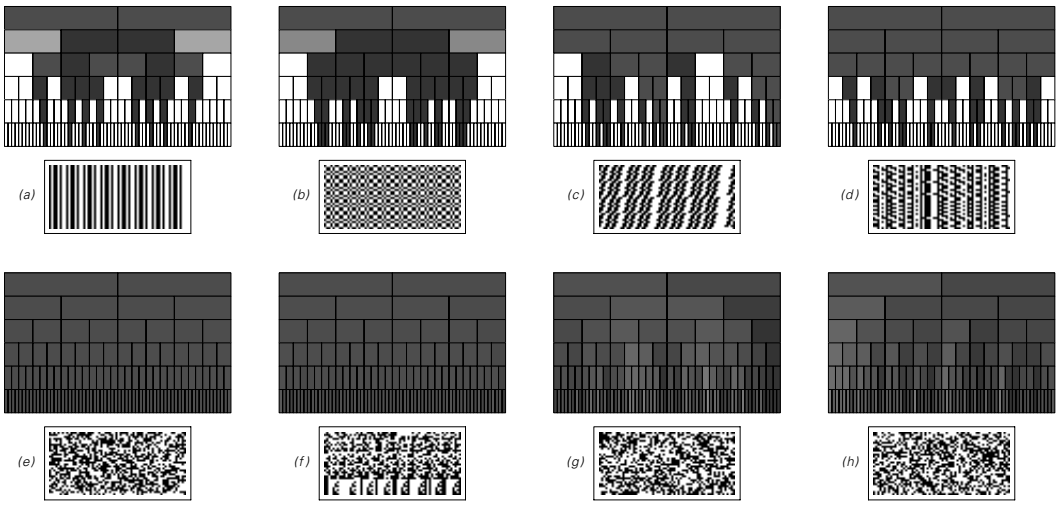
But how can one tell if this is so? What is typically done in practice is to take a sequence that is given and compute from it the values of various specific quantities, and then to compare these values with averages obtained by looking at all possible sequences.

Thus, for example, one might compute the fraction of squares in a given sequence that are black, and compare this to $1/2$. Or one might compute the frequency with which more than two consecutive black squares occur together, and compare this with the value $1/4$ obtained by averaging over all possible sequences.

And if one finds that a value computed from a particular sequence lies close to the average for all possible sequences then one can take this as evidence that the sequence is indeed random. But if one finds that the value lies far from the average then one can take this as evidence that the sequence is not random.

The pictures at the top of the next page show the results of computing the frequencies of different blocks in various sequences, and in each case each successive row shows results for all possible blocks of a given length. The gray levels on every row are set up so that the average of all possible sequences corresponds to the pattern of uniform gray shown below. So any deviation from such uniform gray potentially provides evidence for a deviation from randomness.

And what we see is that in the first three pictures, there are many obvious such deviations, while in the remaining pictures there are no obvious deviations. So from this it is fairly easy to conclude that the first three sequences are definitely not random, while the remaining sequences could still be random.



Statistics of block frequencies for various sequences. In each case the frequency of a particular block is represented by gray level, with results for blocks of successively greater lengths being shown on successive rows as indicated on the left. The original sequences are shown broken into lines and arranged in two dimensions. Sequences (b), (c) and (d) are generated by substitution systems with rules (b) ■ → ■□, □ → □■, (c) ■ → ■■□, □ → □ and (d) ■ → ■■■□□□, □ → ■□ respectively. (Note that these substitution systems are the simplest ones that yield equal frequencies of all blocks up to lengths 1, 2 and 3 respectively.) Sequence (e) is generated by a linear feedback shift register (essentially an additive cellular automaton) with tap positions *{2, 11}*. Sequence (f) is formed by concatenating base 2 digits of successive integers. Sequence (g) is the center column of the pattern generated by the rule 30 cellular automaton. Sequence (h) is the base 2 digits of π.

And indeed sequence (a) is certainly not random; in fact it is purely repetitive. And in general it is fairly easy to see that in any sequence that is purely repetitive there must beyond a certain length be many blocks whose frequencies are far from equal.

It turns out that the same is true for nested sequences. And in the picture above, sequences (b), (c) and (d) are all nested.

But what about the remaining sequences? Sequences (e) and (f) seem to yield frequencies that in every case correspond accurately to those obtained by averaging over all possible sequences. Sequences (g) and (h) yield results that are fairly similar, but exhibit some definite fluctuations.

So do these fluctuations represent evidence that sequences (g) and (h) are not in fact random? If one looks at the set of all possible sequences, one can fairly easily calculate the distribution of frequencies for any particular block. And from this distribution one can tell with

what probability a given deviation from the average should occur for a sequence that is genuinely chosen at random.

The result turns out to be quite consistent with what we see in pictures (g) and (h). But it is far from what we see in pictures (e) and (f). So even though individual block frequencies seem to suggest that sequences (d) and (e) are random, the lack of any spread in these frequencies provides evidence that in fact they are not.

So are sequences (g) and (h) in the end truly random? Just like other sequences discussed in this chapter they are in some sense not, since they can both be generated by simple underlying rules. But what the picture on the facing page demonstrates is that if one just does statistical analysis by computing frequencies of blocks one will see no evidence of any such underlying simplicity.

One might imagine that if one were to compute other quantities one could immediately find such evidence. But it turns out that many of the obvious quantities one might consider computing are in the end equivalent to various combinations of block frequencies. And perhaps as a result of this, it has sometimes been thought that if one could just compute frequencies of blocks of all lengths one would have a kind of universal test for randomness. But sequences like (e) and (f) on the facing page make it clear that this is not the case.

So what kinds of quantities can one in the end use in doing statistical analysis? The answer is that at least in principle one can use any quantity whatsoever, and in particular one can use quantities that arise from any of the processes of perception and analysis that I have discussed so far in this chapter. For in each case all one has to do is to compute the value of a quantity from a particular sequence of data, and then compare this value with what would be obtained by averaging over all possible sequences. In practice, however, the kinds of quantities actually used in statistical analysis of sequences tend to be rather limited. Indeed, beyond block frequencies, the only other ones that are common are those based on correlations, spectra, and occasionally run lengths—all of which we already discussed earlier in this chapter.

Nevertheless, one can in general imagine taking absolutely any process and using it as the basis for statistical analysis. For given some

specific process one can apply it to a piece of raw data, and then see how the results compare with those obtained from all possible sequences.

If the process is sufficiently simple then by using traditional mathematics one can sometimes work out fairly completely what will happen with all possible sequences. But in the vast majority of cases this cannot be done, and so in practice one has no choice but just to compare with results obtained by sampling some fairly limited collection of possible sequences.

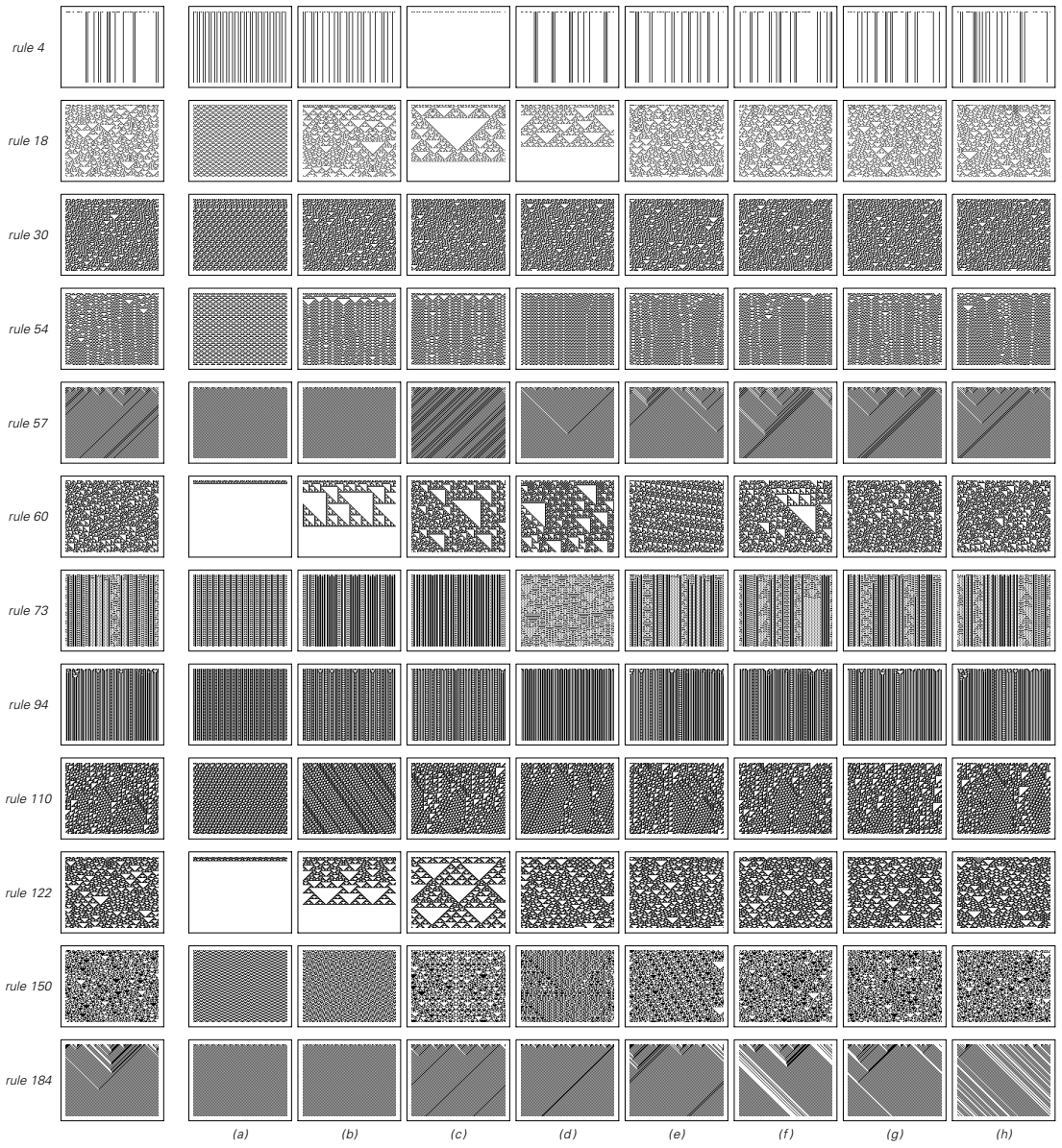
Under these circumstances therefore it becomes quite unrealistic to notice subtle deviations from average behavior. And indeed the only reliable strategy is usually just to look for cases in which there are huge differences between results for particular pieces of data and for typical sequences. For any such differences provide clear evidence that the data cannot in fact be considered random.

As an example of what can happen when simple processes are applied to data, the pictures on the facing page show the results of evolution according to various cellular automaton rules, with initial conditions given by the sequences from page 594. On each row the first picture illustrates the typical behavior of each cellular automaton. And the point is that if the sequences used as initial conditions for the other pictures are to be considered random then the behavior they yield should be similar.

But what we see is that in many cases the behavior actually obtained is dramatically different. And what this means is that in such cases statistical analysis based on simple cellular automata succeeds in recognizing that the sequences are not in fact random.

But what about sequences like (g) and (h)? With these sequences none of the simple cellular automaton rules shown here yield behavior that can readily be distinguished from what is typical. And indeed this is what I have found for all simple cellular automata that I have searched.

So from this we must conclude that—just as with all the other methods of perception and analysis discussed in this chapter—statistical analysis, even with some generalization, cannot readily recognize that sequences like (g) and (h) are anything but completely random—even though at an underlying level these sequences were generated by quite simple rules.



Examples of applying various rules for cellular automaton evolution to the sequences from page 594. The picture at the left-hand end of each row is chosen to show the typical behavior of each cellular automaton, given arbitrary initial conditions. Each cellular automaton rule in effect corresponds to a different statistical analysis procedure. Rule 4 picks out isolated black cells. Rule 60 essentially constructs a difference table for the sequence of elements. Rules 57 and 184 test for the overall density of black cells. (As indicated by page 136 the preponderance of white stripes with rule 184 in case (h) is a fluctuation.)