Inside Averages

From X-ray tomography to Plato's books, mathematicians are uncovering secrets hidden in averages

By IVARS PETERSON

Sometimes, it's hard to keep a secret from a mathematician. Secret data contained in confidential files, for instance, can be uncovered if the right mathematical questions are asked. This possibility threatens the security of many information collections now residing in computer data bases.

Suppose you want to find out someone's salary. Normally, that kind of information can't be extracted from, say, the U.S. Census Bureau or the Internal Revenue Service. But you can ask more general questions like: What is the average salary of all males over 40 who live in Pittsburgh?

"By asking a series of questions like this," says mathematician Ronald L. Graham of AT&T Bell Laboratories in Murray Hill, N.J., "it's often possible to compromise the data base and deduce individual pieces of information."

The following example illustrates the idea: You want to find out Arnold's salary. You know that the average of Arnold's and Bob's salaries is $30,000. The average of Arnold's and Charlie's salaries is $32,000, and the average of Bob's and Charlie's salaries is $22,000. This provides enough information to deduce that Arnold's salary is $40,000.

In larger data bases, the computations are more complicated, but the idea is basically the same (see box). In such problems, "you can get your hands only on various averages or other forms of the data," says Graham. "You can't get your hands on the data directly."

Graham and statistician Persi Diaconis of Stanford University have been systematically investigating this type of reconstruction. The kinds of questions that come up are: When you take averages, have you thrown something out or do you still have all the data, just in a slightly different form?

"Using the theory that Ron and I have built, we can give very nice answers to that kind of question," says Diaconis. "It depends on how many averages you use and which ones they are." Some aspects of their theory appear in the PACIFIC JOURNAL OF MATHEMATICS (Vol. 118, No. 2).

The mathematical work of Graham and Diaconis is related to something called the Radon transform, named for an early-20th-century, French mathematician. Radon's work appears to have been pure mathematics carried out for its own sake. But in that last decade or so, it has played a central role in computerized X-ray tomography.

In tomography, hundreds of needlelike X-ray beams are projected through a slice of the human body. Each beam, as it passes through human tissue and bone, becomes weaker, depending on what it encounters. In mathematical terms, firing an X-ray beam through an object and seeing what happens to the beam is equivalent to finding "the projection of a function along a given line." This mathematical operation is a Radon transform.

The task that researchers face involves finding the tissue density at a particular spot, given the final intensities of the various X-ray beams that have passed through the cross section. This is like computing the original data from a set of averages or performing a Radon transform in reverse.

Radon proved the basic theorems that specify under what conditions a function can be reconstructed from various projections or averages. However, Radon's ideas apply only to "continuous" functions. In tomography, an infinite number of X-ray beams would have to be used to sample the entire cross section before the tissue densities at every spot could be accurately computed.

Because only a finite number of X-ray measurements are actually made, mathematicians have had to work out approximate methods that allow cross sections of human organs to be reconstructed with a minimum of error. "You try to make sure that the error isn't right where the tumor is," says Graham. Further mathematical work is leading to faster algorithms for doing these computations so that X-ray images can be generated almost instantaneously.

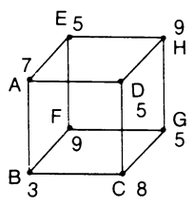
Statisticians are beginning to use a simpler version of the Radon transform idea to look for patterns in complicated arrays of data. In these "projection pursuit" methods, researchers look for special averages (or projections) that in some sense make patterns

A box of averages

In many cases, researchers face a situation in which certain averages are known but the original data are missing. In this simple example, those average values can be identified with the vertices of a cube. Each of the eight numbers shown is the average of its three nearest neighbors. The value at each vertex is "hidden." Thus, the number 7 at vertex A is the average of the hidden values at vertices B, D and E.

These eight averages are enough to reconstruct the original values, says mathematician Ronald L. Graham. In this case, a simple mathematical formula does the job.

For the example shown, the actual value at each vertex is equal to the sum of the nearest-neighbor averages minus double the average at the corner farthest from the point of interest. The actual value at vertex A, for instance, is the sum of the averages at B, D and E minus twice the average at G: $(3+5+5)-2(5) = 3$. The other values are: 6,3,6,9,3,12,9.



"This already points to something that people in tomography noticed in the early days," says Graham. The point that makes the largest contribution is farthest away. Paying scant attention to distant points, he says, has potentially bad implications when you're making approximations in your algorithms. "They can have a bigger effect than you might expect," Graham says.

— Ivars Peterson

more obvious. "Often, you can get a feeling for what's happening in a more complicated situation by scaling it down," says Graham. "You can handle it a little more completely."

A few years ago, Diaconis used projection pursuit methods to determine the order in which the ancient Greek philosopher Plato wrote his most famous books. An account of his work on "projection pursuit for discrete data" is scheduled to appear in the SCANDINAVIAN JOURNAL OF STATISTICS.

This dating question is the kind of thing that scholars have debated for years. Two pieces of information suggest that Plato's books may contain some type of pattern that changes smoothly from his first book to his last.

First, scholars know that *Republic* was written first and *Laws* was written last. Second, Plato, also known as a poet, stated somewhere in his writings that over time, he consciously changed the rhyming pattern in his sentences.

With these clues, more than a decade ago, British researcher L. Brandwood looked at the last five syllables in each of the 3,778 sentences in *Republic* and the endings of sentences in Plato's other books. For each ending, he classified the syllables as either long or short. The result was a lengthy, unwieldy sequence of five-membered sets, ranging from "long-long-long-long-long" through "long-short-short-long-long" and other combinations to "short-short-short-short-short." There were 32 different possible arrangements in all.

Finding a pattern in such a mass of data seemed incredibly difficult. Statisticians, starting with Great Britain's D.R. Cox, tried to fit mathematical models to these data, often assuming that Plato chose sentence endings in the same way that someone would toss a coin to see whether it would come up heads or tails. But it was hard to relate the conclusions drawn from these models to what could be observed in Plato's writing.

Diaconis, on the other hand, decided to analyze the data by taking various averages. This amounted to performing Radon transforms on several data subsets.

Diaconis first looked at the proportion of sentences that ended with a short syllable. Similar proportions were calculated for the other four possible syllable positions. However, both *Republic* and *Laws* showed roughly the same pattern.

Then Diaconis looked at pairs of syllables. "In Plato's *Republic*, a pattern jumped out," he says. In this book, pairs of adjacent syllables tend to be "negatively correlated." If one syllable is short, then the syllable next to it is more likely to be long.

In *Laws*, however, the pattern tends to go the other way. Pairs of adjacent syllables are "positively correlated." Here, on the average, a long syllable is more likely to be followed by another long syllable, while a short syllable is more often followed by another short syllable.

"There's a very clean transition in the rhyming patterns," says Diaconis. "Do pairwise adjacent syllables go together or go opposite? This is a very clean number that I can compute for each book."

Using this special "average," Diaconis worked out an order for Plato's books (see table). This order matched the generally agreed-upon sequence established by Greek scholars using other types of evidence. "Projection pursuit leads to the discovery of striking, easily interpretable structure that does not appear in other analyses of the data," he says.

What the pattern discovered by Diaconis really means in the context of Plato's books isn't clear. For that, a person who knows Greek would have to go back to the books to see what kinds of words are used and how they're placed.

Nevertheless, "it's nice when an abstract math machine makes sense of an age-old question," says Diaconis. "I don't want to claim that I'm solving a big problem, but it does shed some light on a very classical question. Numbers can sometimes do that."

Based on this success and several others, Diaconis has built up a statistical theory that illuminates when "a bunch of averages completely captures the underlying function or data set." His work with Graham involves an algebraic version of the same ideas.

Says Diaconis, "It seems to give a new approach to solving problems that other people have been thinking about for a long time without success." That includes a wide range of mathematical problems such as looking for patterns in ranked data (SN: 3/31/84, p. 202) or determining the security of confidential data bases. □

Analyzing syllable patterns in Plato's books

In the Diaconis analysis, the last five syllables in a given sentence are numbered from one to five. Pairs of adjacent syllables would then be labeled 1,2; 2,3; and so on. For each pair, Diaconis calculated a number that indicates the proportion of sentences with two short syllables in adjacent positions. This number was divided by a factor that takes care of slight differences in the proportion of short syllables in each position.

As shown in the table, numbers less than 1 indicate that a short syllable is more likely to be followed by a long syllable than by a short syllable. Numbers greater than 1 show the opposite correlation. This pattern and some other computed "averages" suggest that Plato wrote his books in the order given (top to bottom), from first to last.

Book	Positions of Adjacent Syllables			
	1,2	2,3	3,4	4,5
Republic	0.88	0.95	0.89	0.91
Timaeus	1.01	0.94	0.99	1.08
Sophist	1.07	1.07	0.97	1.09
Politicus	1.17	1.26	1.05	1.13
Philebus	1.15	1.48	1.02	1.06
Laws	1.07	1.43	1.04	1.02