Computer-Intensive Methods in Statistics

They replace standard assumptions about data with massive calculations. One method, the "bootstrap," has revised many previous estimates of the reliability of scientific inferences

by Persi Diaconis and Bradley Efron

Most statistical methods in common use today were developed between 1800 and 1930, when computation was slow and expensive. Now computation is fast and cheap; the difference is measured in multiples of a million. In the past few years there has been a surge in the development of new statistical theories and methods that take advantage of the high-speed digital computer. The new methods are fantastic computational spendthrifts; they can easily expend a million arithmetic operations on the analysis of 15 data points. The payoff for such intensive computation is freedom from two limiting factors that have dominated statistical theory since its beginnings: the assumption that the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically.

These developments have profound implications throughout science, because statistical theory addresses a grand question: How is one to learn what is true? Suppose 15 measurements of some quantity yield 15 moderately different values. What is the best estimate of the true value? The methods of statistics can answer such a question and can even give a quantitative indication of the estimate's reliability. Because empirical observations are almost always prone to error, conclusions in the sciences (and in many other fields) must often be based on statistical measures of truth. As a result any development that makes statistical inferences more accurate or more versatile can be expected to have broad consequences.

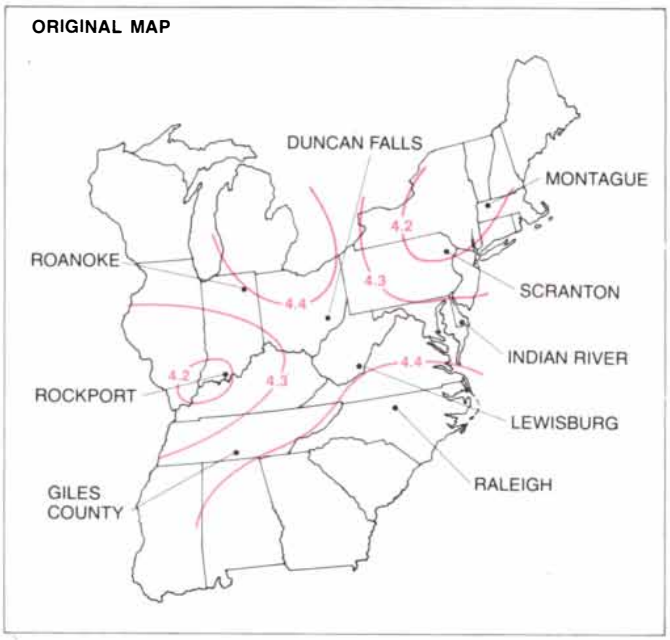
The two advantages of the new methods are best appreciated by comparing them with older ones. First, in older methods it was generally necessary to make certain unverifiable assumptions about the data before statistical analysis could proceed. The assumptions often involved the bell-shaped curve, which is also called the normal or Gaussian distribution, after the German mathematician Carl Friedrich Gauss. When the Gaussian distribution is employed, it is assumed that random fluctuations, or errors, in the experimentally observed values of some quantity are scattered symmetrically about the true value of the quantity. Moreover, it is assumed that the greater the error between the experimental value and the true value is, the less likely it is that the experimental value will be observed. Experience has shown that Gaussian theory works quite well even when the Gaussian distribution is only roughly approximated by the data, which is why statisticians can give reliable predictions even without computers. For sets of data that do not satisfy the Gaussian assumptions, however, the results of statistical methods based on such assumptions are obviously less reliable. Computer-intensive methods can solve most problems without assuming that the data have a Gaussian distribution.

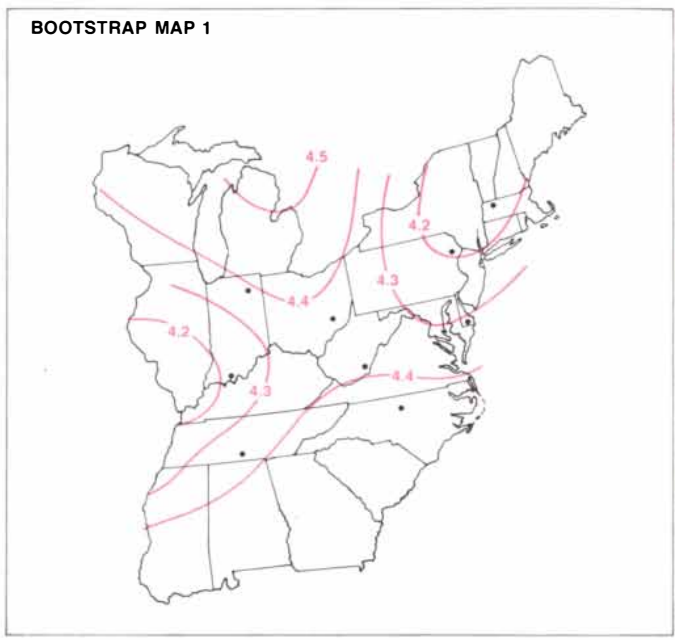
Freedom from the reliance on Gaussian assumptions is a signal development in statistics, but the second advantage of the new techniques is even more liberating. In older practice the arithmetic operations associated with statistical analysis had to be done by hand or with the aid of a desk calculator. Such calculations can often be simplified immensely if the formulas on which the calculations are based have a concise analytical form. Hence statistical theory tended to focus on only a few properties of a statistical sample, such as the mean, the standard deviation and the correlation coefficient, that can easily be manipulated analytically. Many other properties of a sample, however, are of interest to the statistician but are beyond the reach of exact mathematical analysis. The new computer-based methods make it possible to explore such properties numerically, even though their exact analysis is currently impossible. Thus the new methods free the statistician to attack more complicated problems, exploiting a wider array of statistical tools.

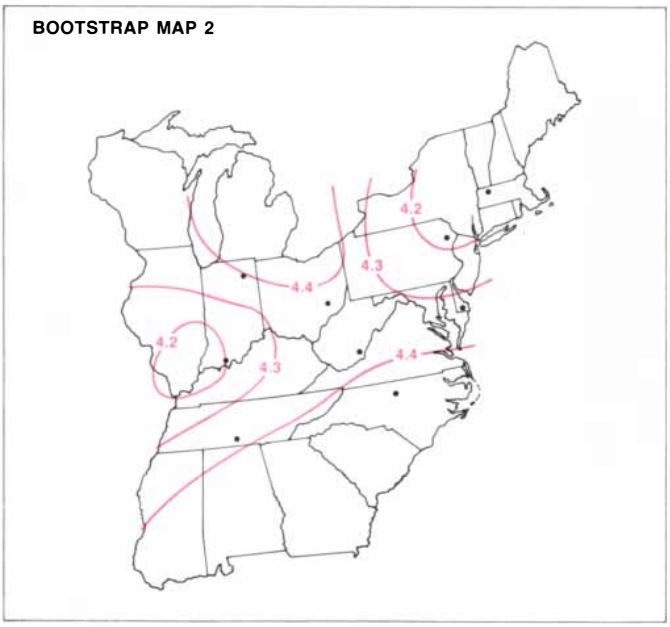
To illustrate how the computer has been applied to statistical inference we have chosen a problem in which only 15 data points appear. We shall apply a method called the bootstrap, invented by one of us (Efron) in 1977, which is quite simple to describe but is so dependent on the computer that it would have been unthinkable 30 years ago.

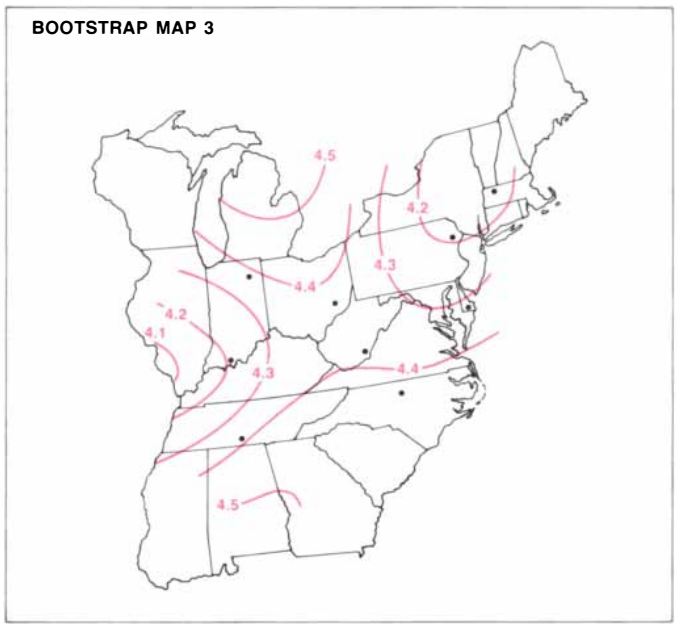
Consider a group of 15 law schools for which two overall characteristics of each entering freshman class are measured: the average undergraduate grade-point average (GPA) and the average

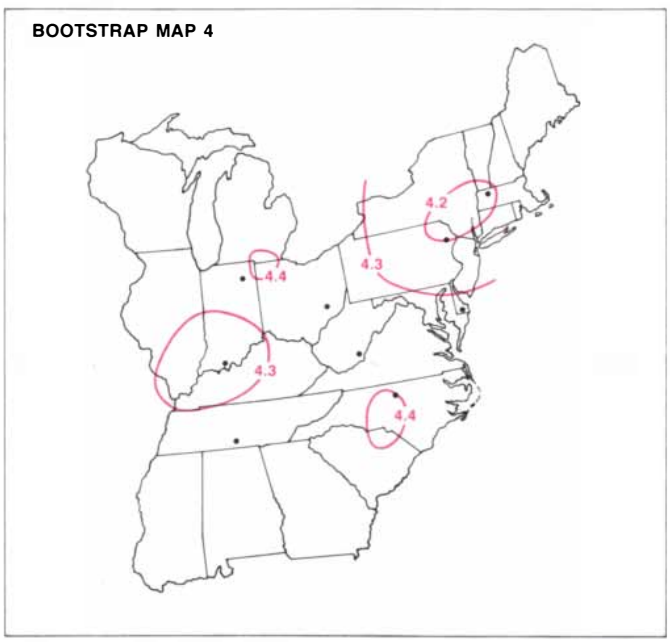
LARGE VARIABILITY of contour lines on a map is revealed by the statistical method called the bootstrap; the method requires so many numerical calculations that it is feasible only with the aid of a computer. The map at the upper left was constructed from 2,000 measurements of the acidity, or *p*H value, of every rainfall recorded at nine weather stations over a period of two years. (The lower the value of the *p*H, the greater the acidity.) The contours were drawn according to a procedure that can be proved optimal under certain conditions. Nevertheless, the 2,000 data points are subject to considerable random variability: contours based on another sample of 2,000 measurements for the same region and time period might have looked quite different. The bootstrap, which was invented by one of the authors (Efron), can estimate from the single set of 2,000 data points the amount of variability the contours would show if many sets of 2,000 data points could be compared. The results of five bootstrap calculations, done with the aid of a computer by Barry P. Eynon and Paul Switzer of Stanford University, are shown in the other five maps. The variability of the contours shows that the original map must be interpreted cautiously: corridors of low acidity on the original map can become islands on subsequent maps.

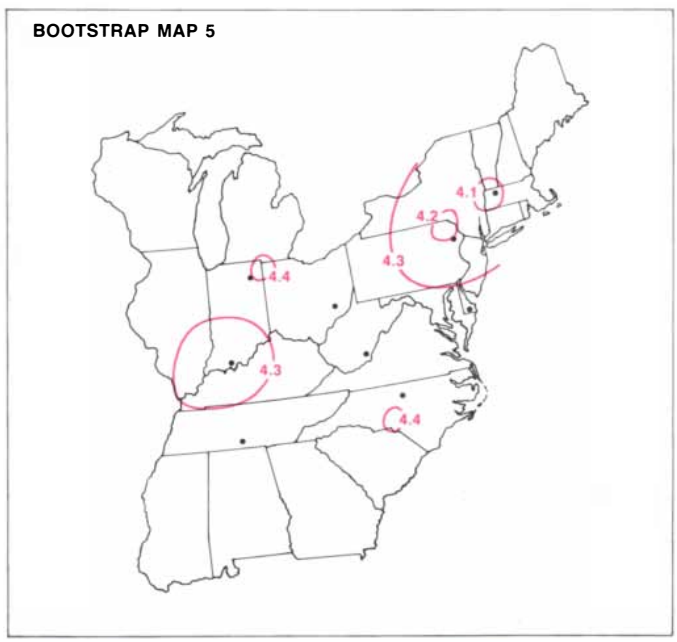


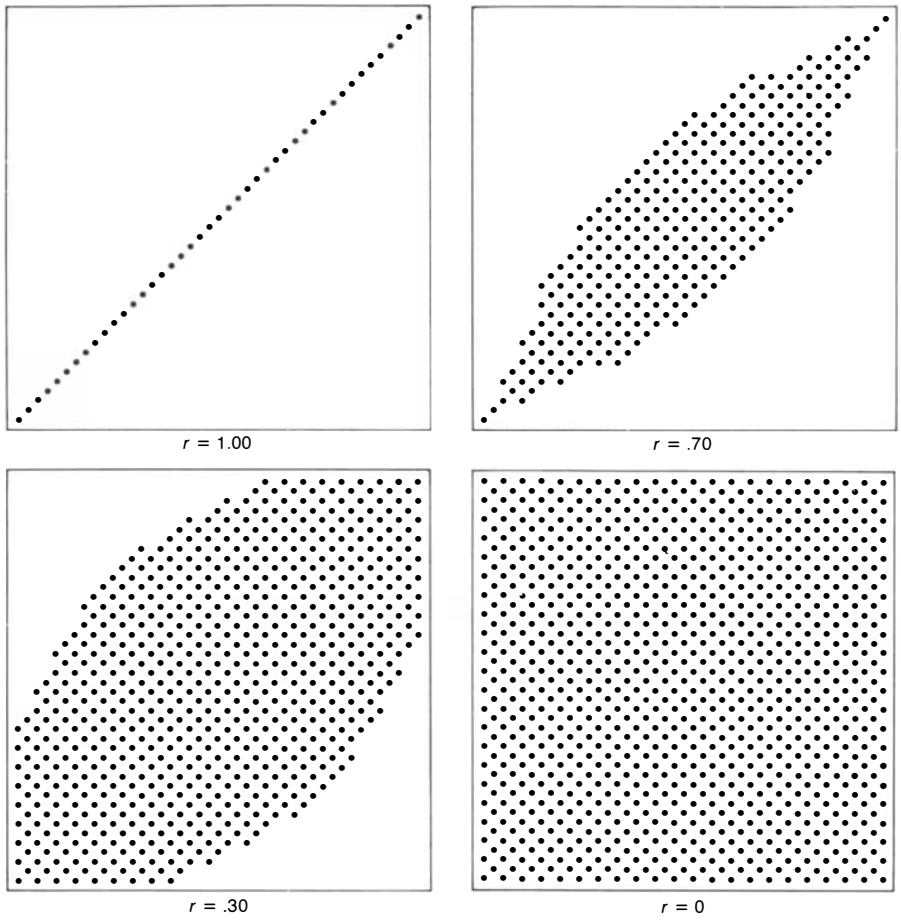




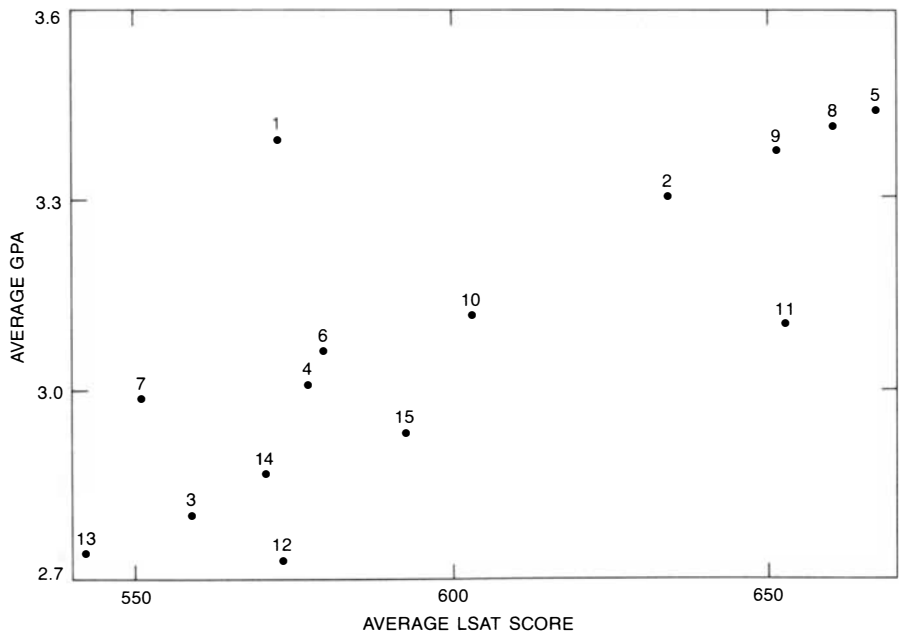








CORRELATION COEFFICIENT is a measure of the tendency of data points plotted on a graph to cluster about a line. The coefficient is usually designated by the letter *r* and can have any value between 1 and −1. The more linear the clustering, the greater the absolute value of *r*.



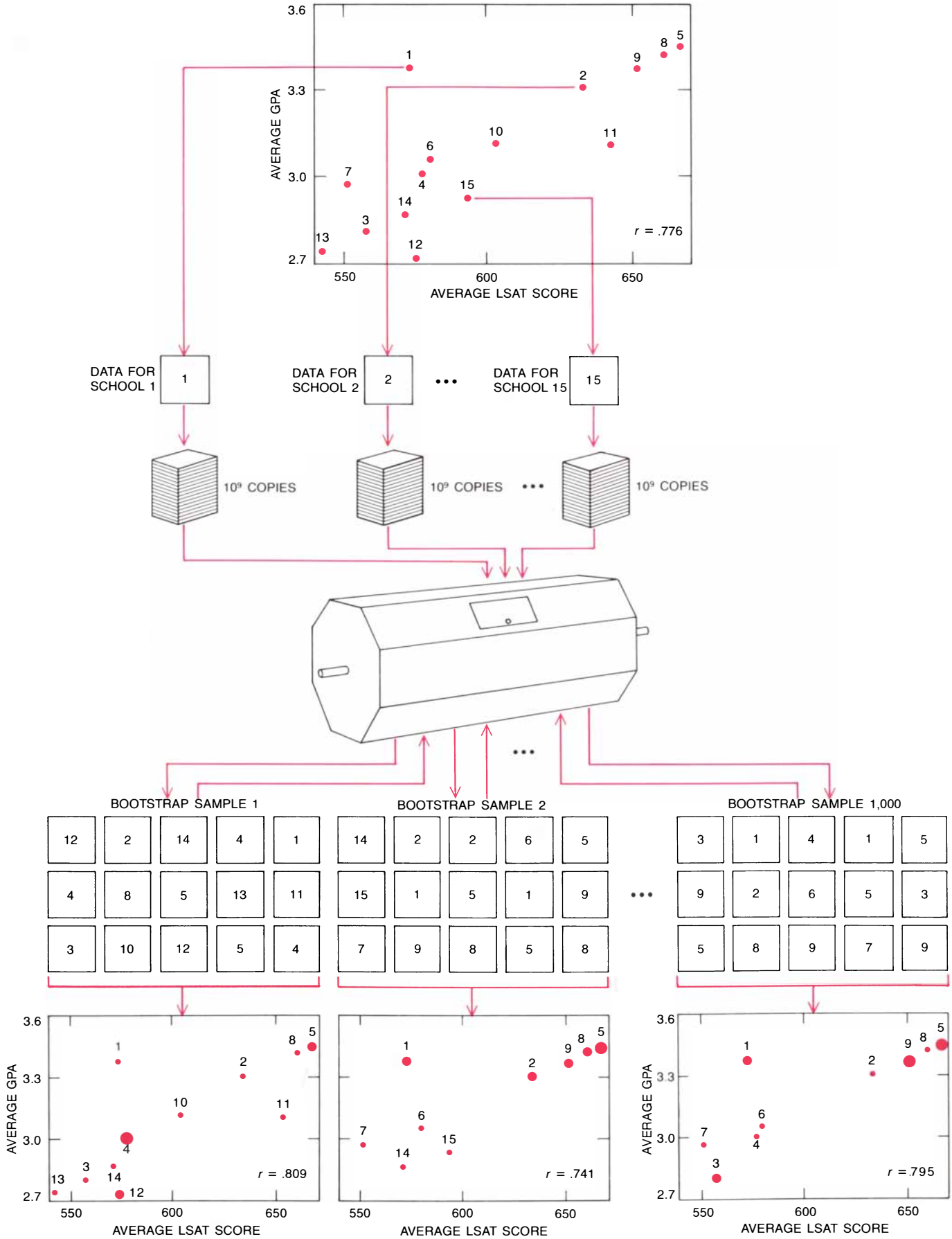
TWO MEASURES for the potential academic performance of the students in the entering classes of 15 American law schools are graphed for each school. Each point on the graph represents the undergraduate grade-point averages (GPA) and the scores on the Law School Admission Test (LSAT), averaged over all the students in one entering class. The graph shows that for the sample of 15 law schools the two measures tend to be proportional: their correlation coefficient *r* is .776. One would like to know how accurately .776 approximates the true value of *r* for all American law schools. That is, one would like to know how much, on the average, the observed value of *r* for a random sample of 15 law schools differs from the true value of *r*.

score on the Law School Admission Test (LSAT). It seems reasonable to suppose the two measures ares roughly proportional to each other: the entering classes with a high average GPA tend to have a high average score on the LSAT. It is unlikely, however, that the proportionality is exact: the entering classes of one or two law schools may show a high average GPA but a low average LSAT, whereas a few other schools may have a low average GPA but a high average LSAT. The statistician wants to know first of all how close the relation between the two measures is to proportionality. Moreover, the statistician must try to estimate the degree to which the available data justify the extrapolation of the first result to all other law schools. In short, how confident can one be that the sample of 15 law schools gives an accurate picture of the population of law schools as a whole?

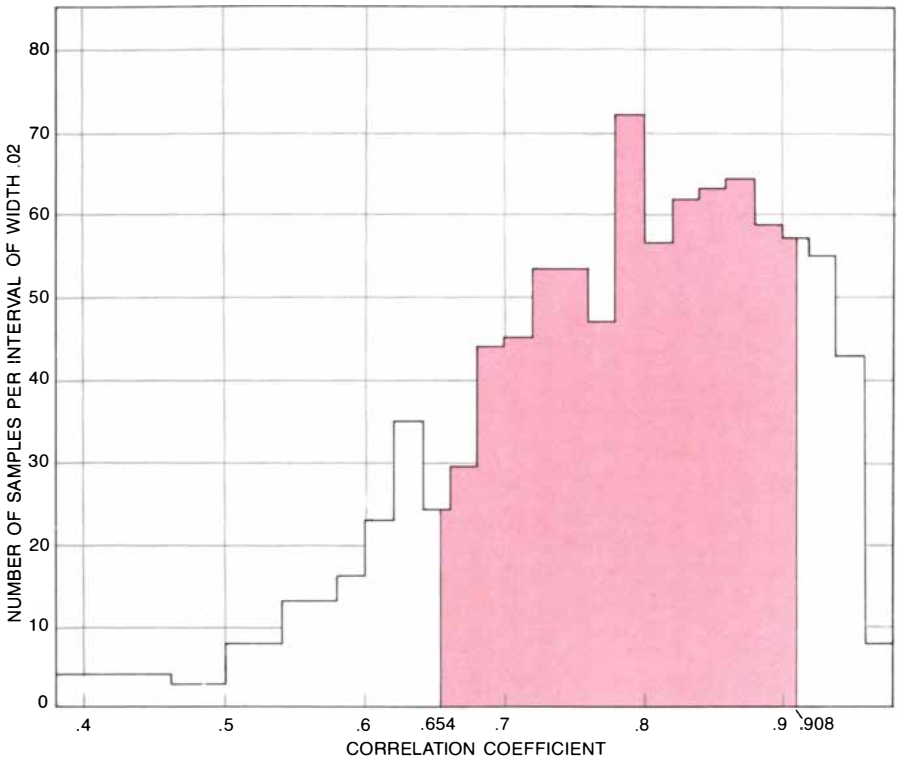
The standard measure of the tendency toward proportionality between two variables such as GPA and LSAT is the correlation coefficient; it is usually designated by the letter *r*. Suppose the data for the law schools are plotted on a graph where the vertical axis represents GPA and the horizontal axis represents LSAT. The correlation coefficient is a measure of the degree to which the points on such a graph tend to cluster along a line. The value of *r* is 0 if the points are scattered at random and gets increasingly close to 1 or −1 as the points tend to cluster along a line of positive or negative slope. (The slope of a line is positive if the line slopes up and to the right, and the slope is negative if the line slopes down and to the right.) The correlation between degrees Fahrenheit and degrees Celsius, for example, is 1 because the two variables are directly proportional to each other. The correlation between the height of fathers and the height of their sons is about .5. Tall fathers tend to have tall sons, but the correspondence is not exact. The correlation between daily consumption of cigarettes and life expectancy has been shown to be negative; that is, the greater the daily consumption of cigarettes, the shorter the life expectancy.

The observed correlation between GPA and LSAT for the 1973 entering classes of the 15 law schools is .776. In other words, there is a strong correlation observed between the two variables and a strong tendency for the points defined by the coordinates of each school to cluster along a line of positive slope. A straightforward mathematical procedure, which takes about five minutes with a desk calculator, was applied to determine the value of *r:* the details of the calculation are not important except that they give a well-defined value of *r* for any collection of data points.

What grounds does one have, how-



BOOTSTRAP METHOD is applied to the sample of 15 law schools shown in the illustration at the bottom of the opposite page in order to assess the accuracy of the correlation coefficient calculated for the sample. The data for each law school are copied perhaps a billion times and all 15 billion copies are thoroughly shuffled. Artificial samples of 15 law schools, called bootstrap samples, are created by selecting sets of 15 data points at random from the 15 billion copies. The value of r is then calculated for each of the bootstrap samples. Although it is simple in concept, the application of the bootstrap requires so many calculations that it is not feasible without a computer.



FREQUENCY DISTRIBUTION of the correlation coefficient *r* is plotted for 1,000 bootstrap samples. A widely accepted measure of the accuracy of a statistical estimator such as *r* is the width of the strip under the central part of its frequency distribution whose area is 68 percent of the area under the entire distribution. The central strip for the bootstrap distribution is shaded in color; its width is .254. Half of the width of the interval, .127, is a good estimate of the average amount by which the observed value of *r* for a sample differs from the true value of *r*.

ever, for believing the true value of *r* is close to .776 for all law schools? After all, the sample could be highly atypical of law schools in general. The law of large numbers guarantees that in large samples the statistical estimate of *r* calculated for the sample is very likely to approach the true value of *r* for the entire population. A sample of only 15 law schools, however, is not a large sample. Hence some measure is needed that can assess the statistical accuracy of the value of *r* given by the sample, namely .776. The bootstrap is intended to provide such a measure.

To understand what statistical accuracy means for an estimate such as *r*, suppose data were available for additional sets of 15 law schools, different from the sets already sampled. For each set of 15 law schools the value of *r* could be calculated, and the amount of variation in the values of *r* for many samples could thereby be described. For example, if 99 percent of the values of *r* calculated for the hypothetical samples were between .775 and .777, one would assign high accuracy to the estimate .776. On the other hand, if the values of *r* were spread out evenly from -1 to 1, the estimate of *r* given by the original sample would have no accuracy and would therefore be useless. In other words, the statistical accuracy of an estimated value of *r* depends on the width of the interval bracketing the estimated value that is associated with a certain percentage of all the samples. Unfortunately the data needed to calculate the value of *r* for many different samples are generally lacking. Thus, because the law school example is intended to reflect the conditions of real statistical practice, we shall assume for the moment that the only data available are those for the original sample of 15 law schools. Indeed, if more data were available, they could be employed to give a better estimate for the value of *r* than .776.

The bootstrap procedure is a means of estimating the statistical accuracy of *r* from the data in a single sample. The idea is to mimic the process of selecting many samples of size 15 in order to find the probability that the values of their correlation coefficients fall within various intervals. The samples are generated from the data in the original sample. The name bootstrap, which is derived from the old saying about pulling yourself up by your own bootstraps, reflects the fact that the one available sample gives rise to many others.

In effect, the bootstrap samples are generated as follows. The data for the first school are copied an enormous number of times, say a billion, and the data for each of the other 14 schools are copied an equal number of times. The resulting 15 billion copies are thoroughly mixed. Samples of size 15 are then selected at random and the correlation coefficient is calculated for each sample. On a computer the steps of copying, mixing and selecting new sets of data are all carried out by a procedure that is much faster but mathematically equivalent: the computer assigns a number to each law school and then generates the samples by matching a string of random numbers to the numbers that correspond to the law schools.

The samples generated in this way are called bootstrap samples. The distribution of the correlation coefficients for the bootstrap samples can be treated as if it were a distribution constructed from real samples: it gives an estimate of the statistical accuracy of the value of *r* that was calculated for the original sample. We generated 1,000 bootstrap samples from the data for the 15 law schools in our sample. Of the 1,000 samples there were 680, or 68 percent, whose correlation coefficients were between .654 and .908. The width of this interval, .254, is the bootstrap measure of accuracy of the value of *r* for the sample. Half of the width of the interval, .127, can be interpreted as the bootstrap estimate of the average amount by which the observed value of *r* for a random sample of size 15 differs from the true value of *r*.

It is worth noting that the statistical accuracy cannot be defined simply as the accuracy of an individual estimate such as .776, that is, the difference between the estimate and the true value of *r*. In a real problem this difference can never be known; if it were known, the problem would vanish, because one could subtract the difference from the estimate and so obtain the true value exactly. Instead the statistical accuracy refers, as we have indicated, to the average magnitude of the deviation of the estimate from the true value.

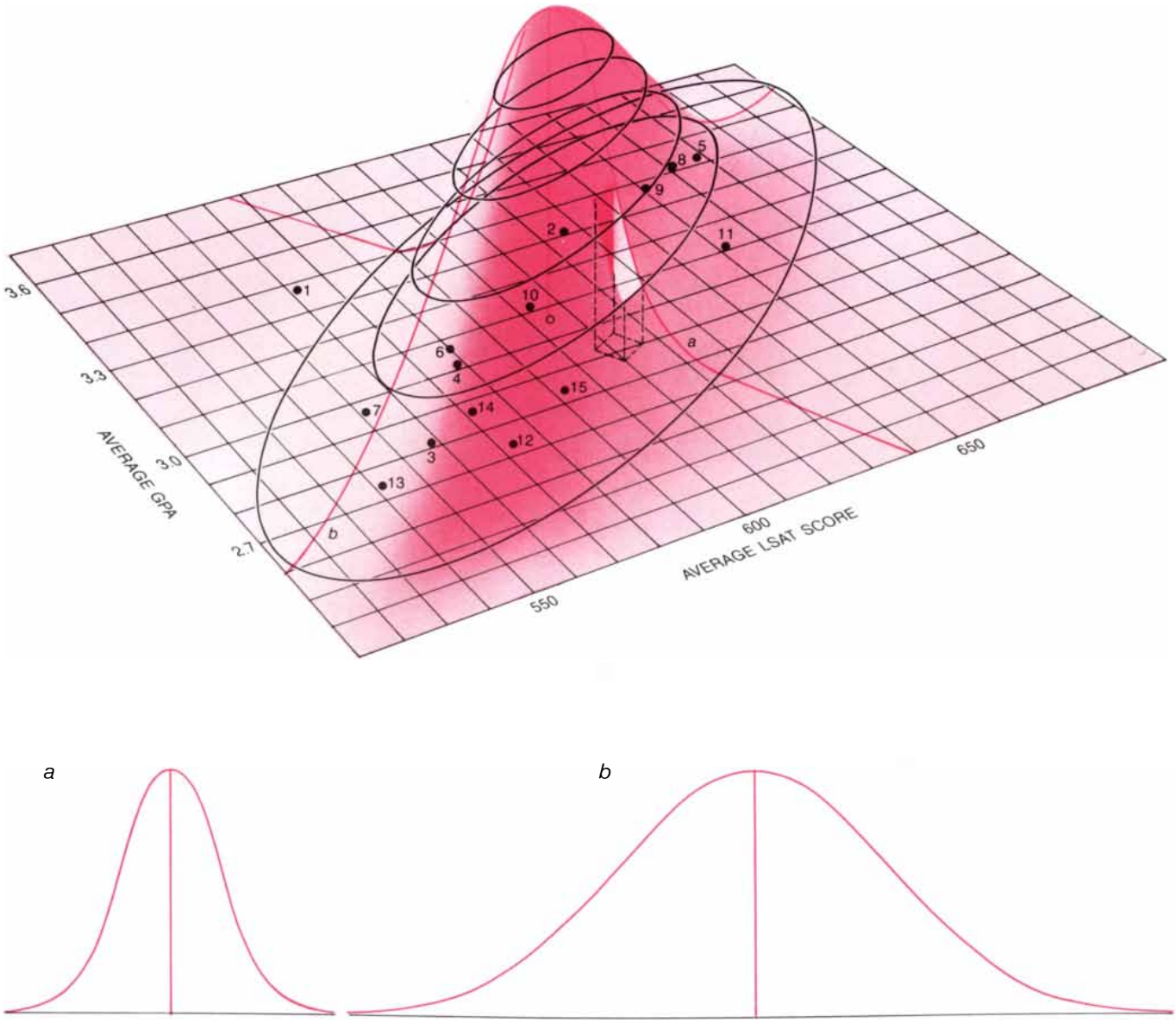
If the results of the bootstrap distribution can be taken as a measure of the statistical accuracy of the original estimate, then .776 is a rough estimate but not entirely worthless. The true correlation coefficient, that is, the value of *r* for the population of law schools as a whole, could well be .6 or .9, but it is almost certainly not zero. Our theoretical work shows that the bootstrap measure of statistical accuracy is dependable in a wide variety of situations.

We can now abandon our self-imposed ignorance because in the law school example the accuracy of the estimated correlation coefficient can be tested directly. Indeed, we chose the example because all the data for average GPA and average LSAT scores of American law school students in 1973

have already been gathered. There were 82 American law schools in 1973, and the correlation between GPA and LSAT for all the schools was .761. (Hence .761 is the true value of r we referred to above, a quantity that would not be known in most situations.) More important, the true statistical accuracy of the estimate given by the original sample can be calculated, because the distribution of the values of r for many real samples of size 15 can be determined. Samples of size 15 can be chosen at random from the 82 law schools in 82^{15}, or about 5×10^{28}, equally likely ways. In principle the value of r could be calculated for each sample, and so the number of samples for which r lies within a small interval could be plotted for intervals of equal size. The resulting graph is called a frequency distribution.

In practice the frequency distribution for samples of size 15 selected from the 82 law schools can only be approximated; a computer set to work at the beginning of the big bang calculating r for each of the 82^{15} samples at a rate of a billion a second would still not have finished the task. Instead r is calculated for a large but manageable number of randomly selected samples of size 15, say a million samples.

We found that 68 percent of the correlation coefficients for a million samples were grouped between .606 and .876, an interval whose width is .270. In other words, if a sample of 15 law schools is selected at random, the probability that its correlation coefficient lies between .606 and .876 is .68. Note that the width of the interval is in good agreement with that defined by 68 percent of



BELL-SHAPED SURFACE was employed in 1915 by Sir Ronald Fisher in his method for estimating from a single sample how much the correlation coefficient varies from sample to sample. In order to make such an estimate by Fisher's method it is necessary to assume that all the data points in the sample are selected according to probabilities given by the bell-shaped surface. The surface is constructed to fit the data in the sample. In the law school example the highest point of the surface must lie directly over the point on the plane where the GPA and the LSAT points both have their overall average values (*open circle*). The slope and orientation of the surface with respect to the plane of the graph depend on how the data points are scattered. The contours of equal height on the surface are elliptical, and the cross sections are bell-shaped curves of varying width; two cross sections are shown in the lower part of the illustration. Fisher's method can be interpreted as choosing bootstrap samples from among all the points on the plane of the graph. The probability of choosing a point from within a given region on the graph is equal to the volume that lies between that region and the bell-shaped surface (*volume of "hole"*) divided by the entire volume that lies between the surface and the graph. By carrying out the bootstrap sampling with a computer only for the discrete points in the original sample the probability distribution that is given by the bell-shaped surface need not be assumed.

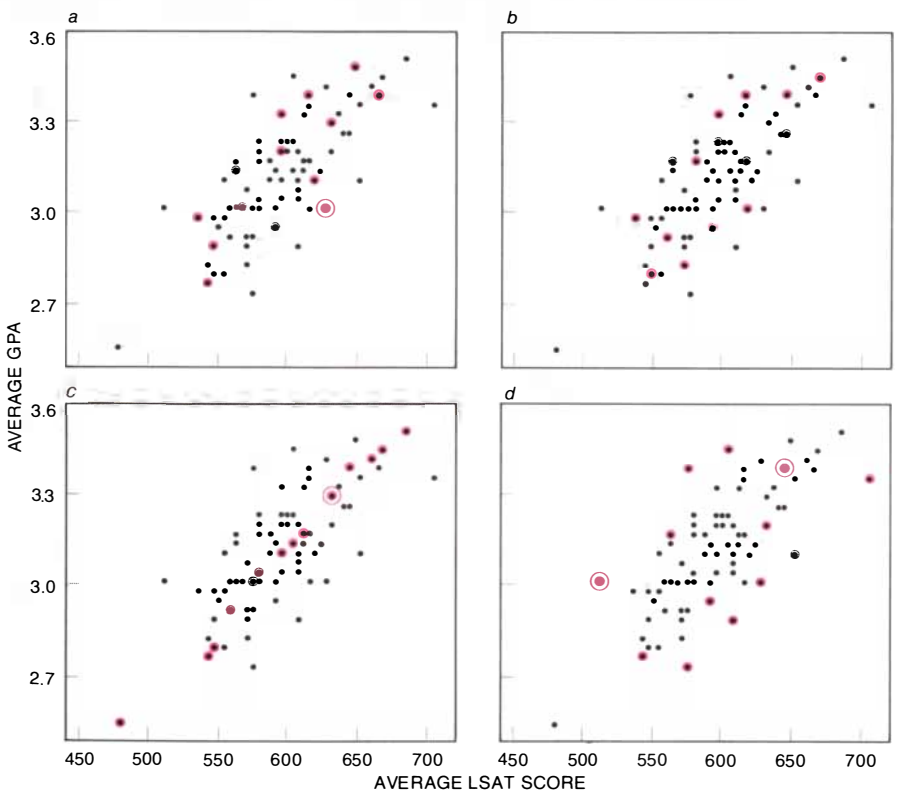
the bootstrap samples, even though the endpoints of the intervals coincide only roughly.

It turns out that the agreement is no accident. Theoretical investigations done by Rudolph J. Beran, Peter J. Bickel and David A. Freedman of the University of California at Berkeley, by Kesar Singh of Rutgers University and by us at Stanford University show that for the correlation coefficient and for a wide variety of other statistics the interval associated with the bootstrap distribution and the interval associated with the distribution of the real samples usually have nearly the same width. (Intervals that include 68 percent of the samples are commonly cited for comparison because for a bell-shaped curve 68 percent of the samples lie within one standard deviation of the peak of the bell.)

At first glance this theoretical result seems paradoxical: it suggests that from the information in each sample one can derive a good approximation to the frequency distribution of the correlation coefficient for all real samples of the same size. It is as if statisticians had discovered the statistical analogue of the hologram, a pattern of light waves that is preserved on a surface. The scene from which the light waves are emitted can be reconstructed in great detail from the whole surface of a hologram, but if pieces of the surface are broken off the entire scene can still be reconstructed from each piece. Not every sample is like a broken piece of a hologram, however; the good properties of the bootstrap are good average properties. Like any other statistical procedure, the bootstrap will give misleading answers for a small percentage of the possible samples.

Suppose the correlation coefficient for the sample of 15 law schools had been nearly 1. That is, suppose all the data points in the sample lay almost perfectly along a straight line. The circumstance is extremely unlikely, given the real data for the 82 law schools, but it could happen. It would then follow that every sample generated by the bootstrap procedure would also lie along the same straight line and so every bootstrap value of r would be nearly equal to 1. The width of the interval associated with 68 percent of the bootstrap samples would therefore be approximately zero. According to the bootstrap procedure, the statistical accuracy of the estimated value of r would be almost perfect, which is incorrect.

The bootstrap does not always guarantee a true picture of the statistical accuracy of a sample estimate. What has been proved is that the bootstrap gives a good picture of the accuracy of the estimate most of the time. There are always a few samples for which the bootstrap does not work, and one cannot know in advance which they are. The limitation is not so much a failure of the bootstrap procedure as it is a restatement of the conditions of uncertainty under which all statistical analyses must proceed.



STATISTICAL ACCURACY of the observed value of r for a random sample can be known precisely only if it is known how r varies for a large number of samples. The 15 law schools for which the value of r has been calculated were selected at random from the total population of 82 American law schools. The data points in each of the four graphs represent the average GPA and average LSAT score for each of the 82 law schools. There are 82^{15} ways to choose samples of 15 law schools from the total population; four such samples have been selected by circling the points in color. (It is possible to select a school more than once in a given sample; such schools have been circled more than once.) The observed values of r for samples a and b are roughly equal to the true correlation coefficient for all 82 schools. The value of r for sample c, however, is much too high and the value of r for sample d is much too low. The true variability in the value of r for samples of 15 law schools can be determined by finding its value for many such samples because data for many more than 15 law schools (in fact, for all 82 of them) are available. Additional data, however, are often impossible to obtain. The bootstrap can estimate the amount of variability that would be shown by all the samples on the basis of one sample.

What are the advantages of applying the bootstrap? In order to appreciate them it is useful to describe how the accuracy of the correlation coefficient (and of most other statistics) was calculated before the computer became widely available. The earlier procedure can be described in terms of the bootstrap, although it goes without saying that before the invention of the computer statisticians did not characterize their methods in such terms. In 1915 the British statistician Sir Ronald Fisher calculated the accuracy of r theoretically. Fisher had to assume that the data for the two variables, average GPA and average LSAT in our example, were drawn at random from a normal probability distribution, represented by a bell-shaped surface. The surface is a two-dimensional analogue of the one-dimensional bell-shaped curve. There is a family of such surfaces whose shape and orientation can be chosen to fit the available set of data. The surface is fitted to the data in the law school sample by placing the top of the bell directly over the point on the graph where both the GPA and the LSAT points have their overall average values. The surface slopes downward to the graph at a rate that depends on how widely the data points are scattered [*see illustration on preceding page*].

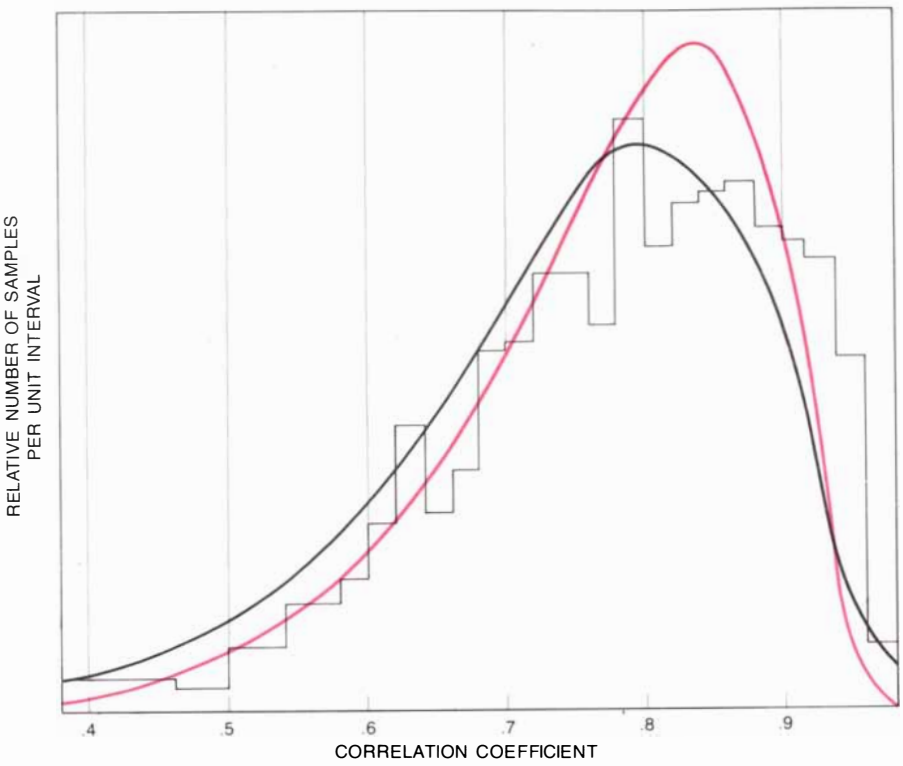
The bell-shaped surface is interpreted as a probability distribution in the same way the graph of values of r for law school samples is a frequency distribution. The probability of selecting a point on the graph of GPA and LSAT scores from within a certain region is equal to the volume that lies under the bell-shaped surface and directly above the region, divided by the entire volume of the space that lies between the surface and the graph. Fisher was then able to

generate a distribution for the values of r by bootstrapping from the bell-shaped probability distribution. In effect, many samples of 15 data points are selected from the graph according to the probability given by their position under the bell-shaped surface. The value of r is calculated for each sample and a frequency distribution for the values of r is plotted. According to Fisher's method, the width of the interval that includes 68 percent of the values of r is .226, in good agreement with the true value of .270 but not as close in this case as the bootstrap estimate of .254.

The bulk of Fisher's calculation can be done analytically because of the assumption that the data in the sample are selected from a normal probability distribution. This assumption is a disadvantage of the method, however, because it might not be true. It is certainly not true in the law school example. Moreover, even if it is true, there is no easy way to check it; in most situations a much larger sample, with perhaps several hundred data points, would be needed to verify the shape of the surface.

The calculations involved in the bootstrap, in which there are no simplifying assumptions about the probability distribution, would have been quite impractical 30 years ago. As we have mentioned, the calculation of a single correlation coefficient takes about five minutes with a desk calculator, and one must carry out from 50 to 1,000 such calculations before a reasonably accurate frequency distribution for the bootstrap samples can be determined.

Today the calculation of a single value of r takes a ten-thousandth of a second with a medium-size computer; at such speed the bootstrap becomes feasible for routine application. If 1,000 bootstrap samples are generated, all the calculations necessary to estimate the width of the interval that includes 68 percent of the samples take less than a second and cost less than a dollar. The cost estimate is based on performing about 100,000 arithmetic operations. More ambitious bootstrap analyses, which give more detailed information about the accuracy of r, require about a million arithmetic operations.



BOOTSTRAP DISTRIBUTION of the correlation coefficient r (*"skyline" in black*) closely approximates the true distribution of r (*smooth curve in black*). The true distribution is actually plotted for a million samples of size 15, chosen at random from the 82^{15} such samples that can be selected from the 82 law schools; differences between the distribution graphed here and the distribution that could in principle be plotted for all 82^{15} samples are not discernible. The shape of the bootstrap distribution also approximates the shape of the distribution that can be estimated according to the probabilities given by a bell-shaped surface (*smooth curve in color*). The agreement suggests the bootstrap can be employed as a measure of the accuracy with which the correlation coefficient of the sample predicts the correlation coefficient of the population. The rather close agreement among the peaks of the distributions is an artifact of the sample.

The bootstrap is not limited to the analysis of the variability of statistics, such as the correlation coefficient, that are mathematically simple. It has been applied to many problems for which the variability of a statistic cannot be assessed analytically. Consider a family of statistics called principal components, which were introduced by Harold Hotelling of Columbia University in 1933. Principal components were devised to solve problems such as the following one, given in a textbook by Kantilal V. Mardia and John T. Kent of the University of Leeds and John M. Bibby of the Open University.

Eighty-eight college students each take two closed-book tests and three open-book tests. Suppose, for the purpose of grading the students, one wants to find the weighted average of the five scores that generates the greatest differences among the students. (In order to make the ratios and not merely the differences of the overall scores as variable as possible, the weights must be scaled so that the sum of their squares is equal to 1.) One set of weights arises if only the final test score is considered; the weights assigned would then be 0, 0, 0, 0 and 1. If all the students had high scores on the final test, however, the summary score generated by this set of weights would not be effective for differentiating the students. Another summary score arises if each test is given equal weight; the weights would then all be equal to $1/\sqrt{5}$, or about .45. The set of weights for the five tests that gives the greatest differences among the students is called the first principal component.

The first principal component is impossible to describe in a mathematically closed form; it must be computed numerically. When the calculation is done for the 88 students, the weights of the first principal component turn out to be roughly equal to one another. The greatest distinctions can therefore be made among the students by finding the average of the five scores.

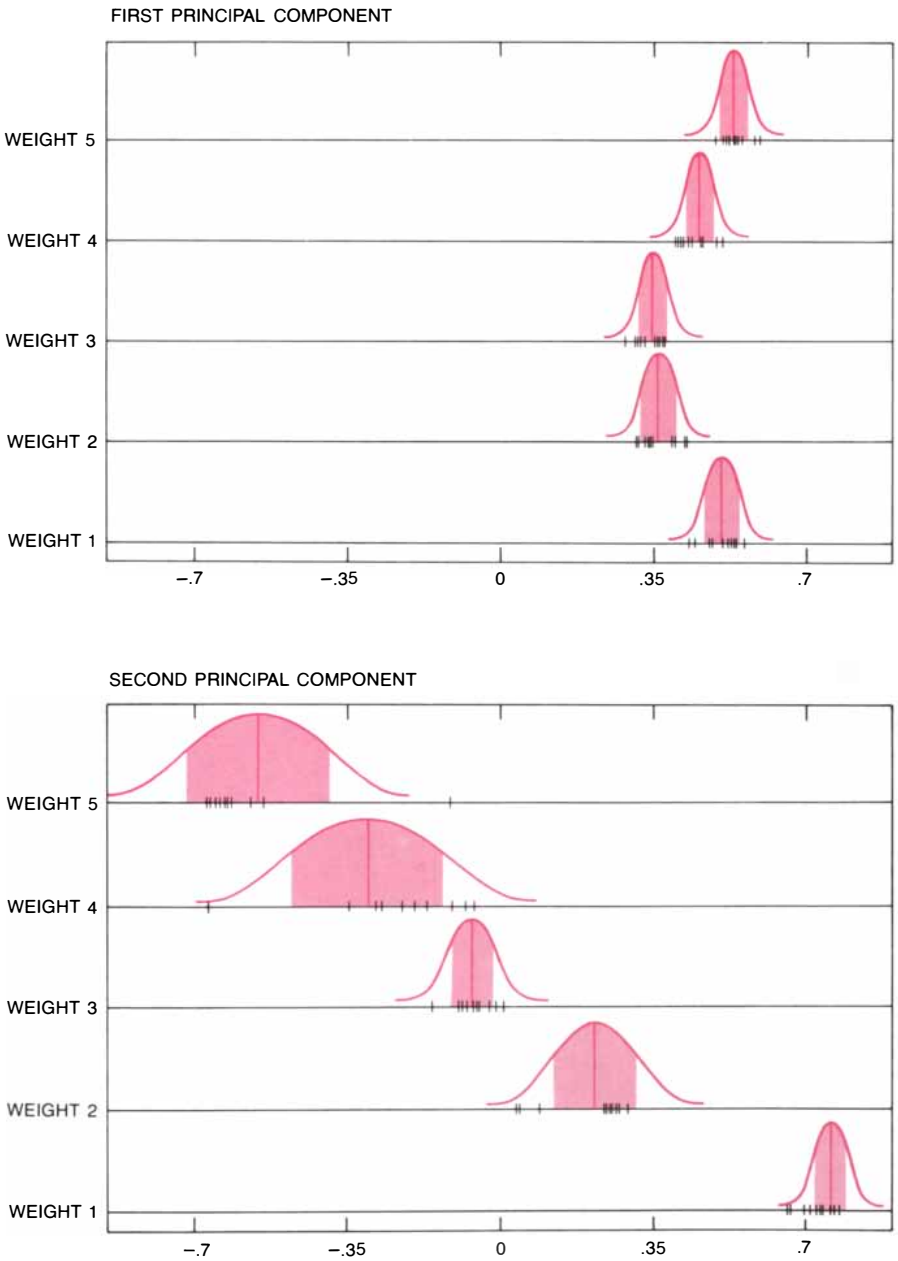
The second principal component is the set of weights, subject to a mathematical constraint of independence, that gives the second-greatest differences among the students. When the second principal component is calculated for the 88 students, the weights turn out to give the difference between an average of the open-book scores and an average of the closed-book scores. The principal components suggest useful and unexpected interpretations for the averages of the student scores. How reliable are the interpretations? If they are to be trusted, one must try to determine how much variation there is in the values of the two principal components for samples of 88 students selected at random.

The problem of quantifying the variability of principal components for samples of a given size has preoccupied many statisticians for the past 50 years. If the appropriate normal distribution is

assumed, partial solutions to questions concerning the frequency distribution for the first principal component can be given; little is known, however, about the second component and higher ones. By applying the bootstrap method a computer can quickly give an estimate of variability for any principal component without assuming that the data have a normal distribution.

In principle the bootstrap analysis is carried out just as it is for the correlation coefficient. Each student's set of five test scores is copied many times (that is, all five scores are written on the same piece of paper) and the copies are thoroughly shuffled. A new sample of size 88 is drawn at random and the principal components are calculated for it. The sampling is repeated many times and a frequency distribution is plotted for each principal component.

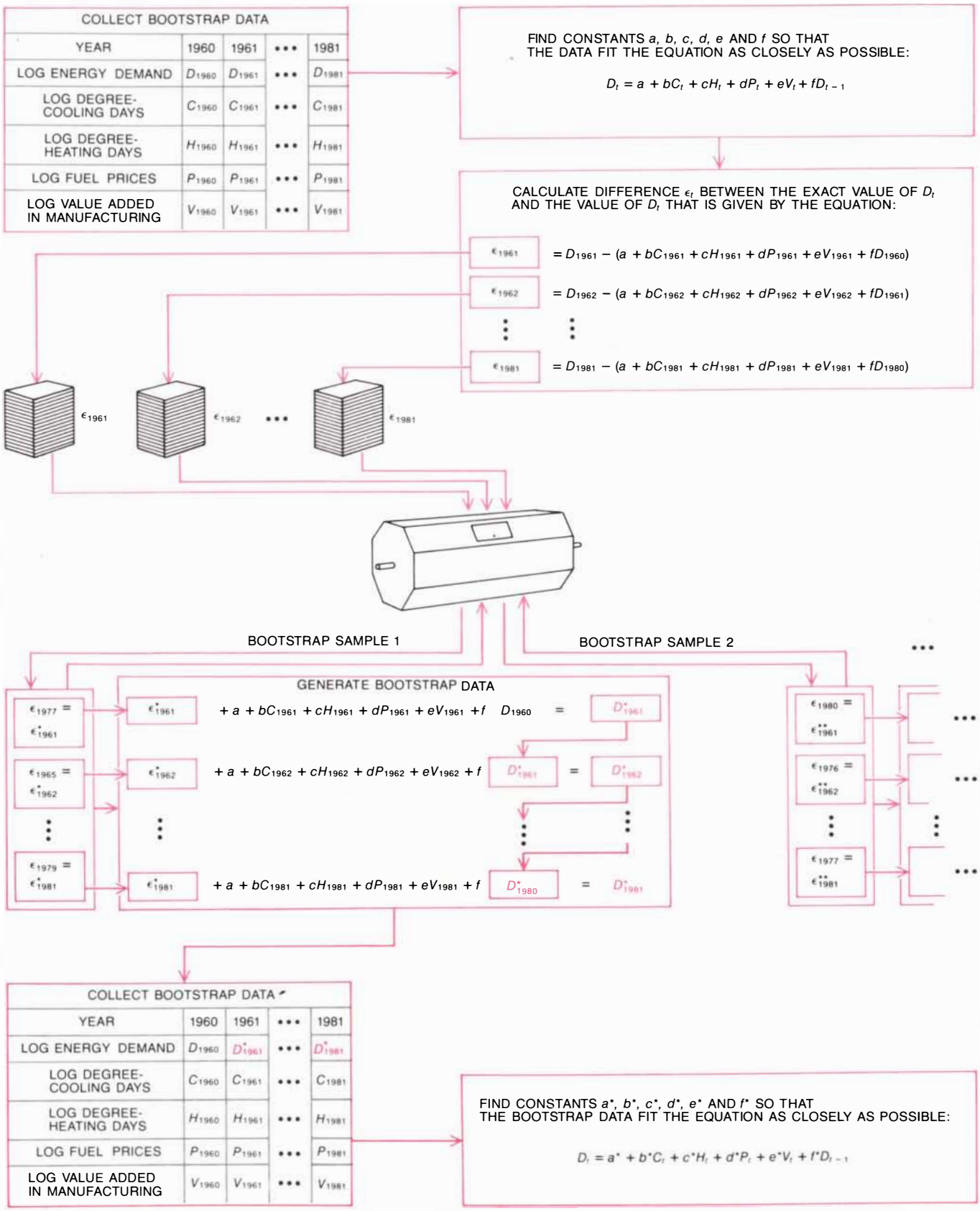
The results suggest that the weights associated with the first principal component are quite stable: they vary only in the second decimal place. The weights associated with the second principal component are less stable, but in a structured way. Remember that the second principal component was interpreted as the difference between an average of the open-book tests and an average of the closed-book tests. The interpretation is confirmed by the bootstrap analysis, but the weights given to the open-book tests are quite variable. The distribution for the principal components generated by the bootstrap is a good estimate of the true distribution of the principal components for samples of size 88. It takes only about two seconds for a large computer to do 100 bootstrap replications.



PRINCIPAL COMPONENTS are statistical estimators that have been widely applied for calculating summary scores on standardized tests. Suppose 88 students take five tests each and suppose, in order to assign a summary score, one wants to find the weighted average of the five test scores that generates the greatest differences among the students. The first principal component is the set of weights that solves the problem. The second principal component is the set of weights, subject to a mathematical constraint of independence, that generates the second most variable combination of the test scores. To assess the variability of the two principal components for many additional samples of 88 students the bootstrap was applied to the single sample. Each student's score for the five tests was written on a piece of paper, and each set of five scores was copied many times. All the copies were then shuffled and bootstrap samples of size 88 were selected at random. The principal components were calculated for each bootstrap sample. The variation in the weights for the first 10 bootstrap samples is shown by the black ticks on each graph; the red vertical lines indicate the observed values of the weights. The width of the central strip under the small bell-shaped curves indicates the variability of the weights. The fourth and fifth weights of the second principal component are particularly unstable.

Not every statistical estimator is a number. Nine weather stations in the eastern and midwestern U.S. recorded the *p*H level, or acidity, of every rainfall from September, 1978, through August, 1980. (A *p*H value of less than 7 is acidic, and the lower the *p*H value, the greater the acidity.) During the two years 2,000 *p*H values were measured. To represent the data Barry P. Eynon and Paul Switzer of Stanford prepared a *p*H-contour map of the region; the *p*H values are constant along a contour line. Such a map can be generated from the data by a well-defined mathematical procedure called Kriging, after the South African mining engineer H. G. Krige. Although the contour map is strictly determined by the data, it represents an extrapolation from the data collected at nine stations to many points in space and time (in fact, to an infinite number) that are not included in the original sample. One can therefore ask how variable the contours on the map would be if random variations yielded different samples of 2,000 *p*H values.

In this example neither the true contour map nor the true variability of all contour maps generated by samples of 2,000 *p*H values can be known. Both estimates must be made from the original data alone if they are to be made at all. By bootstrapping the original sample of 2,000 *p*H values in a way that preserves the geographic relations among the weather stations, Eynon and Switzer generated the maps shown in the illustration on page 117. There is no generally accepted measure of the variability of contour lines on a map, analogous to the width of an interval of a frequency distribution. Intuitively, however, the variability is readily perceived. It shows that the original contour map must be inter-



MODEL OF ENERGY DEMAND called RDFOR (Regional Demand Forecasting Model) was employed by the U.S. Department of Energy to analyze and forecast energy demand in 10 regions of the U.S. For each region the data are fitted as closely as possible to a mathematical model called a regression equation. The demand for energy in any given year is assumed to depend on the demand for energy the year before as well as on several other measures. Each error term ϵ_t is the difference between the predicted value of the energy demand in a given year and the observed value. Bootstrap samples of the error terms are selected at random and artificial data for energy demand are generated by the method shown in the diagram. The bootstrapped data are then fitted to a new regression equation, and the variability of the regression equations generated by the bootstrap gives an estimate of the expected accuracy of the model in predicting energy demand. A bootstrap analysis done by David A. Freedman of the University of California at Berkeley and Stephen C. Peters of Stanford has shown that the variability of the regression equations is from two to three times greater than was previously thought.

preted cautiously. Corridors of relatively low or relatively high acidity on the original map can shrink to become islands on a bootstrap map, depending on the effects of random "noise."

Statistical estimation is often carried out by making the available data conform as closely as possible to some predetermined form, or model. The simplest models are the line, the plane and the higher-dimensional analogues of the plane. Consider the graph of the 15 data points that represent the 15 law schools. Intuitively there are many lines that could be drawn to represent the trend of the data points, and so it is reasonable to agree in advance on a precise method for fitting the points to a line. Probably the most widely used estimator in statistics is a method for fitting points to a line called the least-squares method. The method was invented by Gauss and by Joseph Louis Lagrange in the early 19th century in order to make astronomical predictions.

The least-squares line is the line that minimizes the sum of the squares of the vertical distances between the data points and the line. A straightforward calculation gives the equation of the least-squares line from the data points. If the bootstrap is applied to the data, fake data sets can be generated, and the least-squares method can be applied to each fake set of data points in order to fit them to a new line. The fluctuation of the lines generated by the bootstrap shows the variability of the least-squares method as a statistical estimator for this set of data points.

The least-squares method and its generalizations are particularly useful for complex problems in which an investigator must bring large amounts of diverse information to bear on a single question. The U.S. Department of Energy, for example, has developed a model called the Regional Demand Forecasting Model (RDFOR), which attempts to forecast the demand for energy in 10 regions of the U.S. In the model it is assumed that the energy demand for each region in a given year depends in a simple way on five variables: the amount of variation above 75 degrees F. in summer, the variation below 65 degrees in winter, the price of fuel, the value added in manufacturing (a measure of the economic conditions in the region) and the energy demand during the previous year.

The five variables can be thought of as if they were plotted on a five-dimensional graph, which is exactly analogous to a two-dimensional graph; every point on a five-dimensional graph corresponds to a possible combination of the five variables. The energy demand in a given year associated with a known combination of variables can then be represented by the height of a point in a six-dimensional space above the corresponding point on the five-dimensional graph. The representation of the data in a six-dimensional space is analogous to representing the dependence of some quantity on two other variables as the height of a point in three-dimensional space above a two-dimensional graph. Thus the energy data determine a set of points at various heights in the six-dimensional space.

The least-squares method specifies a way of drawing a five-dimensional analogue of a plane (called a hyperplane) as close as possible to all the points. Because of the dependence of the energy demand on the demand in previous years, the variables must be fitted to the hyperplane by a generalized version of the least-squares method. The generalized method calls for minimizing a weighted sum of errors after the weights have been estimated from the data. In recent years an elaborate method of estimating the accuracy of the procedure and the accuracy of its forecasts has been developed.

Freedman and Stephen C. Peters of Stanford examined the conventional estimates of the accuracy of the procedure by applying the bootstrap. In their approach it is assumed that the data lie close to an appropriate hyperplane, but it is not assumed that the errors between the data points and points that lie on the hyperplane are independent of one another. Instead the relation of the errors from point to point is allowed to have a complicated structure. The bootstrapping of the data was done in a way that

	PATIENT NUMBER						
	149	150	151	152	153	154	155
AGE	20	36	46	44	61	53	43
SEX	M	M	M	M	M	F	M
PRESENCE OF STEROID?	YES	NO	NO	NO	YES	YES	NO
ANTIVIRAL ADMINISTERED?	NO	NO	NO	NO	NO	NO	NO
FATIGUE?	NO	NO	YES	YES	YES	YES	YES
MALAISE?	NO	NO	YES	NO	YES	NO	NO
ANOREXIA?	NO	NO	YES	NO	NO	NO	NO
LARGE LIVER?	NO	NO	NO	NO	YES	NO	NO
FIRM LIVER?	*	NO	NO	YES	YES	NO	NO
PALPABLE SPLEEN?	NO	NO	NO	NO	NO	YES	YES
PRESENCE OF SPIDERS?	NO	NO	YES	NO	YES	YES	YES
PRESENCE OF ASCITES?	NO	NO	YES	NO	NO	NO	YES
PRESENCE OF VARICES?	NO	NO	YES	NO	NO	YES	NO
CONCENTRATION OF BILIRUBIN	.9	.6	7.6	.9	.8	1.5	1.2
CONCENTRATION OF ALKALINE PHOSPHATASE	89	120	*	126	95	84	100
CONCENTRATION OF SERUM GLUTAMIC-OXALOACETIC TRANSAMINASE (SGOT)	152	30	242	142	20	19	19
CONCENTRATION OF ALBUMIN	4.0	4.0	3.3	4.3	4.1	4.1	3.1
CONCENTRATION OF PROTEIN	*	*	50	*	*	48	42
PHYSICIAN'S PROGNOSIS	LIVE	LIVE	*	LIVE	LIVE	*	LIVE
OUTCOME	LIVE	LIVE	DIE	LIVE	LIVE	LIVE	DIE

MEDICAL DATA for seven out of 155 people with acute or chronic hepatitis give the values of 19 variables for each person that, taken together, could predict whether a patient will die or recover from the disease. (An asterisk indicates that information is missing.) It is common practice in statistics to inspect such data before a formal model is constructed; the aim of the inspection is to rule out all but four or five of the most important variables. Peter B. Gregory of the Stanford University School of Medicine eliminated all the variables except the patient's malaise, the presence of ascites (fluid in the abdominal cavity), the concentration of bilirubin and the physician's prognosis. Gregory developed a model based on the four variables that correctly predicted whether or not the patient would die from the disease in 84 percent of the cases.

Grand Touring. What does it mean?

In the new 1983 Toyota Cressida, Grand Touring means Grand Performance. With an amazing new 2.8 liter Twin Cam engine that does to smoothness and quietness of ride, what Michelangelo did to a ceiling. Add that to Cressida's new independent rear suspension, with coil springs and stabilizer bar, and its new electronically controlled 4-speed automatic overdrive transmission and you start to see how grand a driving experience can be.

Cressida's Grand Touring also means Grand Luxury. From the look of luxury outside — dashing, elegant, refined — to the feel of luxury inside — power windows and door locks. Automatic temperature control air conditioning. Variable assist power steering. Cruise control. AM/FM/MPX stereo receiver with four speakers. And a driver's seat that adjusts to your body in four distinct ways.

What else does Grand Touring mean?

In the new Cressida, it represents a feeling you get, while touring town or country. A feeling based on uncompromised performance. And spirited drive. It's the image of the car you've chosen. And of yourself.

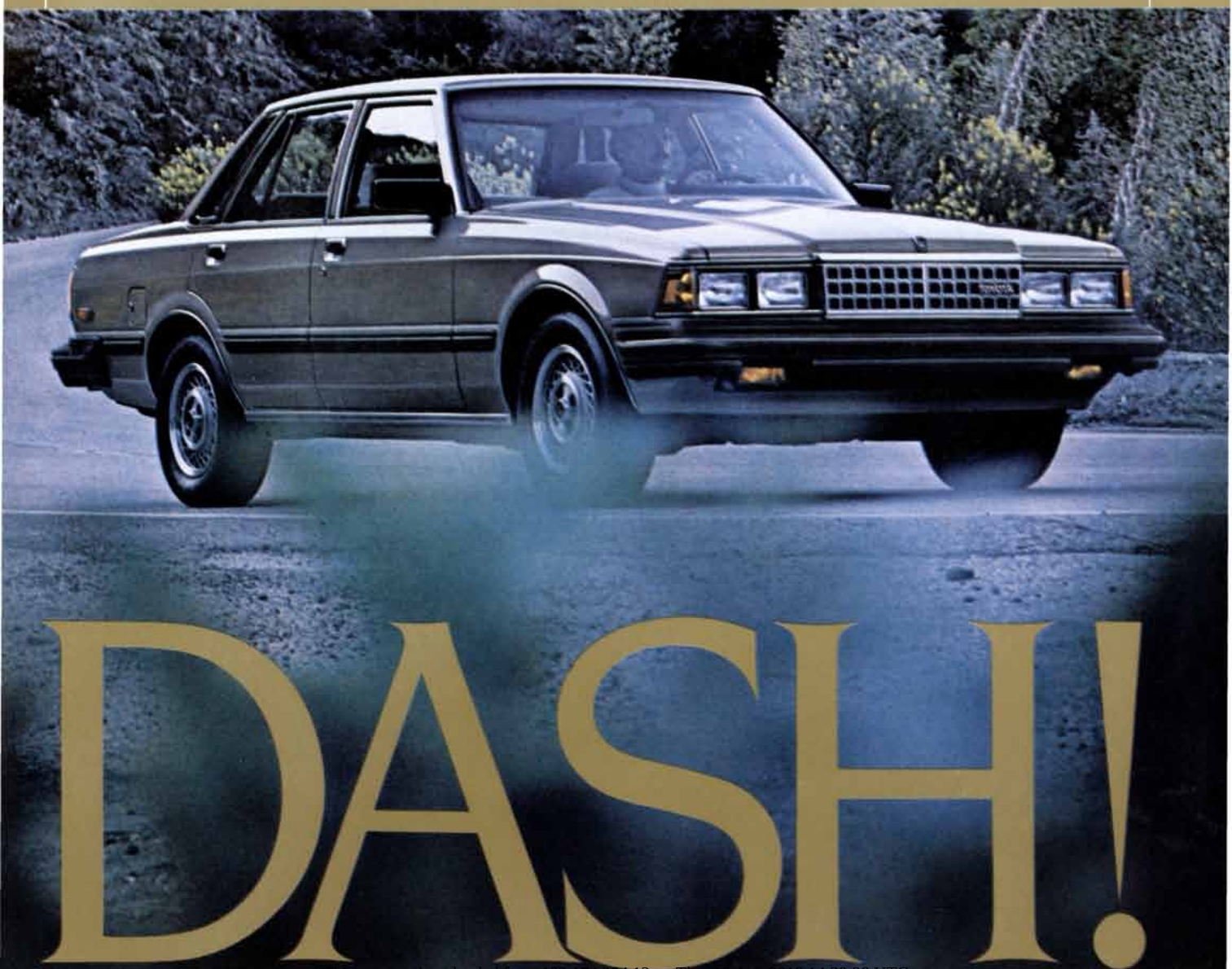
Dashing.

BUCKLE UP—IT'S A GOOD FEELING!

OH WHAT A FEELING!

TOYOTA

CRESSIDA MOVES PROUDLY INTO THE GRAND TOURING CLASS.



preserves the evolution of the energy demand from year to year. The variability of the hyperplanes generated by the bootstrap showed that the standard error previously assumed for the energy model is usually too small by a factor of two or three. The predictions of energy demand made by this aspect of the RDFOR model are therefore much less reliable than was once thought.

The examples we have presented so far have involved clearly defined statistical properties of samples. In practice the data can be inspected, sorted, graphed and preanalyzed in several ways before they are formally analyzed. Estimates of variability that do not take such informal practices into account may not give an accurate picture of statistical variability.

Consider a group of 155 people with acute and chronic hepatitis, initially studied by Peter B. Gregory of the Stanford University School of Medicine. Of the 155 patients 33 died and 122 survived, and for each patient there were data for 19 variables, such as age, sex and the results of standard biochemical measurements. Gregory's aim was to discover whether the data could be combined in a model that could predict a patient's chance of survival.

The analysis of the data was done in several steps. First, all but four of the most important variables were eliminated, because statistical experience suggests it is unwise to fit a model that depends on 19 variables with only 155 data points available. The elimination of the variables was done in two stages: each variable was inspected separately, whereupon six variables that appeared unrelated to the patients' survival were eliminated. A standard statistical procedure was then carried out on the remaining 13 variables, which further reduced the number to four. The variables that remained were the patient's malaise, ascites (the presence of a fluid in the abdomen), the concentration of bilirubin in the liver and the physician's prognosis for the patient. The variables were then fitted to a curve that predicts how the proportion of surviving patients depends on the values of the variables.

Such analysis is typical of scientific practice. In order to estimate its overall variability Gail Gong of Carnegie-Mellon University carried out the entire procedure from preliminary screening through the final curve fitting on bootstrapped samples of the original 155 data points. Her results were surprising and informative. The set of "important variables" generated during the initial stages of the analysis was itself quite erratic. For some bootstrap samples only the prognosis of the physician was found to be important, whereas for others such variables as sex, age, level of fatigue, concentration of albumin and concentration of protein were retained. No single variable emerged as significant in as many as 60 percent of the bootstrap samples.

Although the fitted curve is intended to predict whether or not a patient will survive, it misclassifies 16 percent of the original 155 patients. The estimate of 16 percent, however, is too small because the data on which it is based were also employed to generate the curve. The analysis generated by the bootstrap suggests a better estimate for the probability that the fitted curve will misclassify a given patient is .20.

The prospect of bootstrapping the entire process of data analysis offers hope that an extremely difficult problem will begin to yield, namely the connection between the mathematical theory that underlies statistics and actual statistical practice. The effects of preliminary "data snooping" on the final results are usually ignored, for no better reason than that it is impossible to analyze them mathematically. It now appears that the bootstrap, applied with the aid of the computer, can begin to estimate such effects.

The bootstrap is by no means the only statistical method that relies on the power of the computer. Several other methods such as the jackknife, cross-validation and balanced repeated replications are similar in spirit to the bootstrap but quite different in detail. Each of these procedures generates fake data sets from the original data and assesses the actual variability of a statistic from its variability over all the sets of fake data. The methods differ from the bootstrap and from one another in the way the fake data sets are generated.

The first such method was the jackknife, invented in 1949 by Maurice Quenouille and developed in the 1950's by John W. Tukey of Princeton University and the Bell Laboratories; it has been extensively investigated by Colin L. Mallows of Bell Laboratories, Louis Jaeckel of Berkeley, David V. Hinkley of the University of Texas at Austin, Rupert G. Miller of Stanford, William R. Schucany of Southern Methodist University and many others. The name jackknife was coined by Tukey to suggest that the method is an all-purpose statistical tool.

The jackknife proceeds by removing one observation at a time from the original data and recalculating the statistic of interest for each of the resulting truncated data sets. The variability of the statistic across all the truncated data sets can then be described. For the data from the 15 law schools the jackknife assesses the statistical accuracy of the value of r by making 15 recalculations of r, one for every possible subsample of size 14. The jackknife calls for fewer calculations than the bootstrap but it also seems less flexible and at times less dependable.

Cross-validation is an elaboration of a simple idea. The data are split in half and the second half is set aside; curves are fitted to the first half and then tested one by one for the best fit to the second half. The final testing is the cross-validation; it gives a reliable indication of how well the fitted curve would predict the values of new data. There is nothing special about half splits; for example, the data can be split in the ratio 90 to 10 as

BOOTSTRAP SAMPLE NUMBER	VARIABLES SELECTED
491	ALBUMIN, PROGNOSIS, SEX
492	ASCITES, BILIRUBIN, PROGNOSIS
493	BILIRUBIN, ASCITES
494	BILIRUBIN, PROGNOSIS, MALAISE
495	ASCITES
496	BILIRUBIN
497	ASCITES, VARICES
498	SPIDERS, PROGNOSIS, ALBUMIN
499	AGE, PROGNOSIS, BILIRUBIN, MALAISE, PROTEIN, SPIDERS
500	ASCITES, PROGNOSIS, BILIRUBIN, PROTEIN

VARIABLES DESIGNATED IMPORTANT by informal analysis prior to the construction of a formal statistical model can show wide variation. In a bootstrap study that simulated both the formal and the informal aspects of the statistical analysis, Gail Gong of Carnegie-Mellon University programmed a computer to copy the set of 19 variables associated with each patient many times. The sets of data were thoroughly shuffled and bootstrap samples of 155 sets of data were drawn at random from the collection. Formal and informal techniques of data analysis were then applied to each bootstrap sample, just as they had been for the original sample. The variables chosen as important are shown for 10 of the 500 bootstrap samples generated. Of the four variables originally chosen not one was selected in as many as 60 percent of the samples. Hence the variables identified in the original analysis cannot be taken very seriously.

Exxon has the information processor...

that really can be the start of something big.

At Exxon Office Systems, we're bringing the high-tech office down to earth, by designing office automation to grow the way you grow.

You can start an office automation system simply with just one of our remarkable EXXON 500 Series Information Processors.

Then as you grow, we can grow with you, right into our shared resource office automation system ... The EXXON 8400 Series System.

As a fully functioned system its capability is awesome. Operators can create, edit, reformat, file, share and retrieve documents, all with the touch of a key.

There's a dictionary, an electronic mailbox, a program for keeping calendars and scheduling meetings, a tickler file and more.

Start automating your office now with one of our EXXON 500 Series Information Processors. And step into the future without the shock of unnecessary costs from expanding inefficiently.

For more information on the EXXON 500 or our new office automation system, just send in the coupon below. Or call **800-327-6666**.

EXXON OFFICE SYSTEMS

The future...without the shock.

SAM 05 83

Exxon Office Systems
P.O. Box 10184, Stamford, CT 06904

I'd like to know more about

☐ the EXXON 8400 Series System
☐ the EXXON 500 Series Information Processor

Please have your representative call.

Name _____ Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Telephone _____

800-327-6666.
IN CONNECTICUT, 800-942-2525.

well. Moreover, there is no reason to carry out the cross-validation only once. The data can be randomly split many times in many ways.

Cross-validation has been widely applied to situations in which a curve-fitting procedure is well defined except in some crucial respect. For example, one might be willing to fit a polynomial to the data by the least-squares method, but the degree, or highest power, of the polynomial to be fitted might still be in doubt. (The higher the degree of the polynomial, the less smooth the fitted curve.) Given that half of the data have been fitted by polynomials of various degree, cross-validation can choose the degree of the polynomial that best fits the second half of the data. Seymour Geisser of the University of Minnesota, Mervyn Stone of the University of London and Grace G. Wahba of the University of Wisconsin at Madison have been pioneers in this development.

Instead of splitting the data in half at random a more systematic system of splits can be employed. The splits can be chosen in such a way that the results are optimal for certain simple situations that allow full theoretical analysis. The balanced repeated-replication method, developed by Philip J. McCarthy of Cornell University, makes splits in the data systematically in order to assess the variability of surveys and census samples. Random subsampling, a related method developed by John A. Hartigan of Yale University, is designed to yield dependable confidence intervals in certain situations.

There are close theoretical connections among the methods. One line of thinking develops them all, as well as several others, from the bootstrap. Hence one must ask what assurance can be given that the bootstrap will work most of the time, and how much it can be generalized. To the first question the answer is simple. The bootstrap has been tried on a large number of problems such as the law school problem for which the correct answer is known. The estimate it gives is a good one for such problems, and it can be mathematically proved to work for similar problems.

We have suggested the answer to the second question through the diversity of complex problems to which the bootstrap has already been applied. What is needed for many of them, however, is independent theoretical justification that the bootstrap estimate of accuracy remains as valid as it is for simpler problems. Current theoretical work seeks to provide such justification and to give more precise statements of accuracy based on the bootstrap. Fisher was able to provide a statistical theory that took full advantage of the computational facilities of the 1920's. The goal now is to do the same for the 1980's.

6

Sixth in a series of how Delco Electronics and Bose technology contribute to your enjoyment of driving.

Deceptive... isn't it.

The control panel of this Electronically Tuned Receiver (ETR) is simple—and deceptive. Simple so that the receiver is easy to operate. Deceptive because a very sophisticated technology lies behind it. A technology that produces high fidelity reception from the Delco-GM/Bose Music System under conditions that are even difficult for ordinary radio reception.

The key to this technology is Delco Electronic's own custom integrated circuits. These circuits respond automatically to changing reception conditions and program requirements. So you can enjoy music and driving more.

When you visit your GM dealer* you will understand why Len Feldman wrote in Popular Science: "It's as good as or better than the best home systems I've heard."

*Available as a factory-installed option on Cadillac Seville and Eldorado, Buick Riviera, Oldsmobile Toronado, and Corvette by Chevrolet.

Delco GM BOSE

Sound so real it will change how you feel about driving.