COMPUTER INTENSIVE METHODS IN STATISTICS

BY

PERSI DIACONIS and BRADLEY EFRON

TECHNICAL REPORT NO. 83
JANUARY 1983

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-08

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

COMPUTER INTENSIVE METHODS IN STATISTICS

BY

PERSI DIACONIS and BRADLEY EFRON

TECHNICAL REPORT NO. 83
January 1983

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-08

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also prepared under the Auspices of National Science Foundation Grant MCS80-24649 and issued as Technical Report No. 196, Stanford University, Department of Statistics.

COMPUTER INTENSIVE METHODS IN STATISTICS

Persi Diaconis and Bradley Efron

Abstract

Invited by Scientific American, this paper is intended for a general audience. It explains new developments in theoretical statistics relating to the computer. Most of the discussion concerns various applications of the bootstrap. The applications include a principle components analysis, contour map drawing, an econometric model, and a medical prediction problem.

COMPUTER INTENSIVE METHODS IN STATISTICS

Persi Diaconis and Bradley Efron

New statistical theories are being developed which take advantage of the high-speed computer. *The bootstrap, one such theory, replaces the Gaussian assumptions of classical statistics with massive amounts of computation*.

Most of the commonly used statistical methods were developed between 1820 and 1930, when computation was slow and expensive. Now computation is fast and cheap. The difference is measured in multiples of a million. During the past few years there has been a surge of development in new statistical theories and methods which take advantage of high speed computation. The new methods are fantastic computational spendthrifts by the standards of the past. As we shall see they can easily expend a million arithmetic operations on the analysis of fifteen data points. The payoff for all this computation is freedom from two limiting factors which have dominated statistical theory since Gauss' time (1800), the bell-shaped curve and linear mathematics.

We will focus on one of these new methods, the *bootstrap*. However our main point, that the computer is qualitatively increasing the power of statistical methodology, could be made equally well in a half dozen other contexts discussed in the references: robust regression, recursive partitioning, censored data analysis, projection pursuit modelling, cross-validation, and interactive statistical graphics. These exotic names refer to useful data-analytic devices that may soon sound as familiar to statistical practitioners as correlation and linear regression.

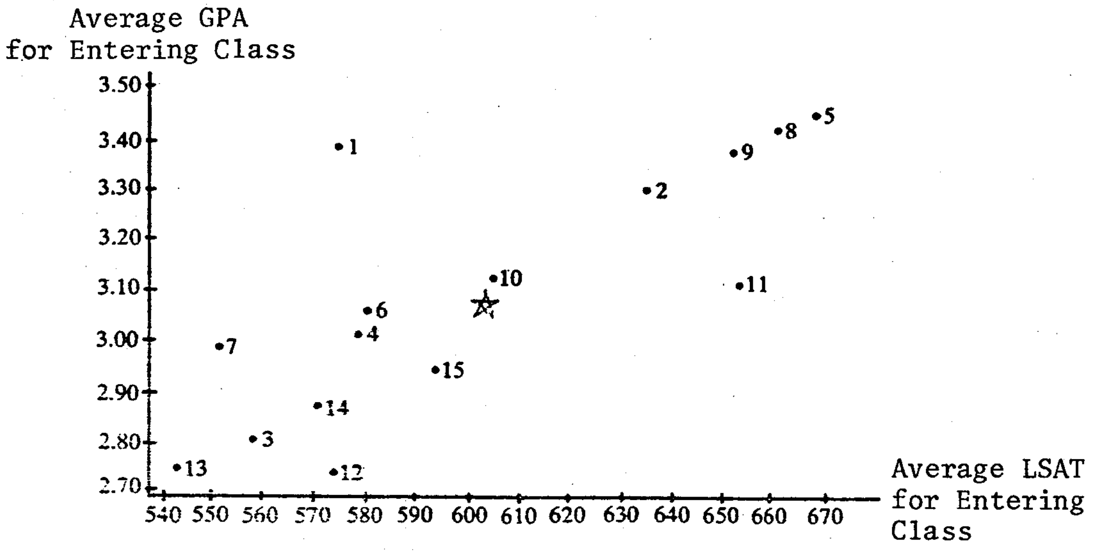
The bootstrap, invented by the second author in 1977, is a simple and widely applicable idea that is particularly dependent on high-speed computation. It can be used to assign a reliable measure of statistical accuracy, the familiar "±"

quantity following a statistical estimate, without concern for the mathematical complexity of the estimator or the non-bell-shapedness of the data. We begin with a small example, involving only a few data points, which nevertheless would have been computationally unthinkable thirty years ago.

* * *

The figure on page A describes the 1973 entering classes at fifteen American law schools. Each point represents one school. The horizontal axis represents the average LSAT score for the entering class (LSAT is a national test for prospective law students), while the vertical axis represents the average undergraduate GPA (grade point average). For example the entering class at the school labelled "1" had average LSAT 576 and average GPA 3.39. The two measures tend to agree with each other: entering classes with high average LSAT tend to have high average GPA, and vice-versa.

The usual statistical measure of this tendency toward agreement is the correlation coefficient "r". This is a numerical measure of agreement between two variables like GPA and LSAT, or between incidence of cancer and level of pollution, or between father's height and son's height. The correlation coefficient r is scaled so the highest possible value is r equals 1, attained when the points lie exactly on a straight line of positive slope - so high values of one variable correspond to high values of the second variable. The lowest possible value is r equals minus 1 which is attained when the points lie on a straight line of negative slope - so high values of one variable correspond to low values of the second variable. For reference, the correlation between fathers heights and sons heights is r equals ½. Tall fathers tend to have tall sons but the relationship isn't perfect. A plot of 15 father-son heights would usually show less agreement than the law school data on page A. The mathematical formula for r, which won't be of concern here, appears below the figure on page A.



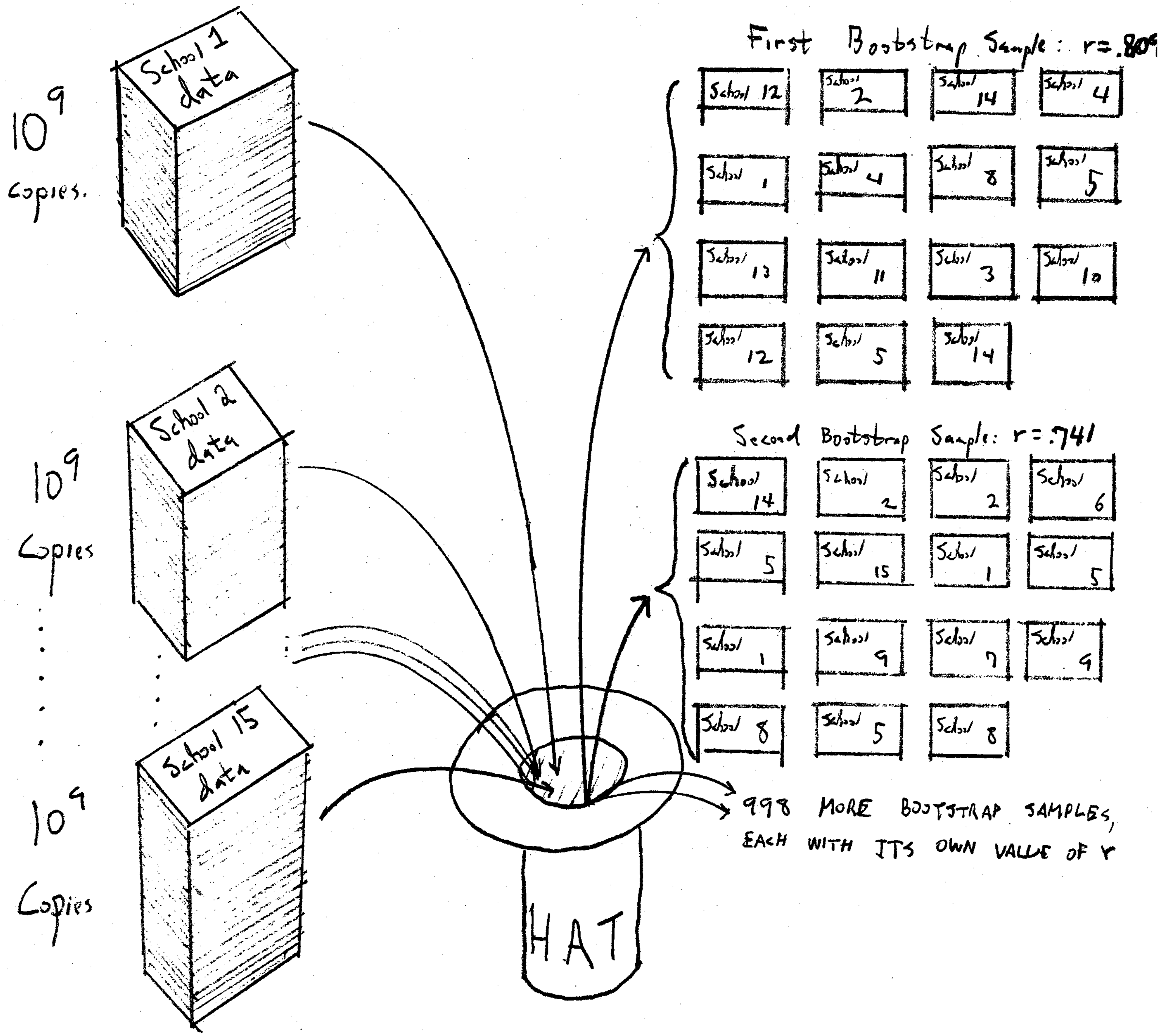
The entering classes of 1973 at 15 American law schools are described by their average LSAT ("law board") scores and their average GPA (undergraduate grade point average). The figure shows that the two measures tend to agree with each other. They have a high correlation coefficient, $r = .776$, compared to $r = 0$ if the two measures were totally unrelated and $r = 1$ if all 15 points lay on a perfect straight line with positive slope. We want to know the accuracy of the estimate .776. The red star is located at the average of the 15 averages. In order to calculate r we need to compute 5 sums, each over the 15 schools, S_1 = Sum of LSAT average scores, S_2 = Sum GPA, S_3 = Sum LSAT^2, S_4 = Sum GPA^2 , and S_5 = Sum LSAT • GPA. The formula for r is $[S_5-S_1 \cdot S_2/15]/[\sqrt{S_3-S^2/15}\ \sqrt{S_4-S_2^2/15}]$. This takes about 5 minutes on an old-fashioned desk calculator, and about 1/10,000 of a second on a computer.

For the 15 law schools, the correlation between LSAT and GPA is calculated to be r equals .776 , indicating a strong positive agreement. Here, r is an estimated correlation coefficient, based on a random sample of size 15, not the true correlation for the entire population of law schools.

How accurate is the estimate r = .776? Is it possible that the true correlation, for the entire population of law schools, is zero, and that the points only seem to show agreement because of random fluctuations? After all, the sample size is only 15 so randomness could conceivably play a major role. The bootstrap is designed to answer just such questions.

To understand what "accuracy" means for an estimate like r , suppose that we had available data from another 15 law schools, different than the ones we actually have, and then another 15, and another 15 etc. etc. For each set of 15 we could compute r , and then we could compare directly how much the r's varied from each other. If all of these imaginary r's lay between .775 and .777 we would assign high accuracy to our actual estimate .776. At the other extreme the values might be spread out evenly from -1 (perfect negative relationship) to +1 , in which case .776 would have no accuracy and be a useless estimate.

Of course we don't really have these hypothetical other sets of 15 law schools. If we did, we would have used them to get a better estimate than .776. The bootstrap algorithm overcomes this difficulty by constructing a sequence of fake data sets using only the data from the original 15 law schools. The construction is illustrated on page B. The data for law school 1, "LSAT = 576, GPA = 3.39", is copied on say a billion slips of paper, and likewise the data for each of the other 14 schools. All 15 billion slips of paper are placed in a hat and well mixed. Then samples of size 15 are drawn from the hat. These are the fake data sets or the "bootstrap samples". For each bootstrap sample we calculate the correlation coefficient. The variability of these bootstrap estimates indicates the accuracy of the actual estimate .776.



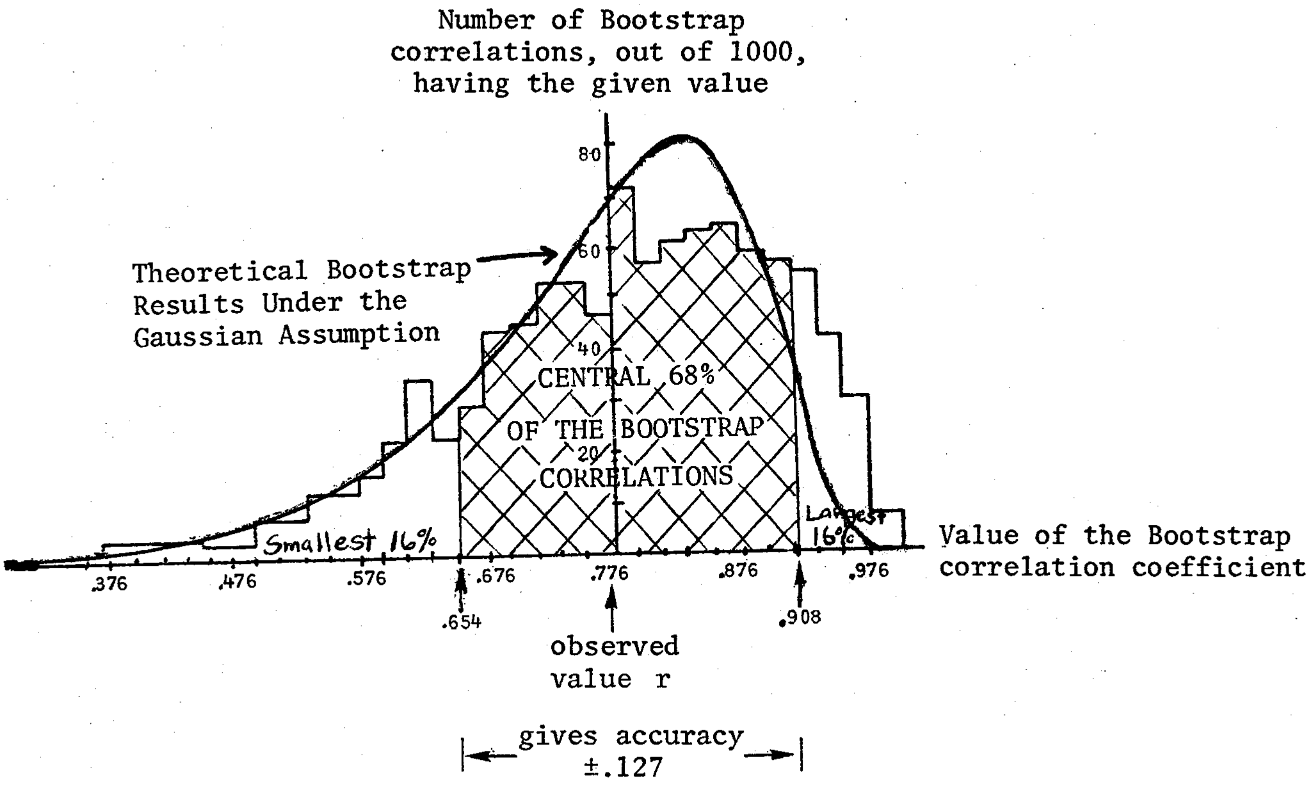
The bootstrap algorithm constructs fake data sets of 15 "new" law schools from the original data. The data for law school 1 is copied some enormous number of times, say one billion, and likewise for law school 2, law school 3, ..., law school 15. All 15 billion copies are placed in a hat and well mixed. Then samples of size 15 are drawn from the hat. These are the fake data sets, or the "bootstrap samples". For each bootstrap sample we recalculate the statistic of interest, in this case the correlation coefficient. The variation of these bootstrap statistics is a dependable measure of accuracy for the actual statistic.

One thousand bootstrap samples, each of size 15, were "drawn from the hat" for the law school data. (The bootstrap sampling procedure is actually carried out on the computer, using a random number generator to make the selections rather than multitudinous slips of paper in a hat.) The resulting 1000 bootstrap correlation coefficients are shown in the figure on page C. The central 680 of them, 68%, lay between .654 and .908. Half the length of this interval gives ±.127 as an accuracy measure for the observed value $r = .776$. This particular way of interpreting the bootstrap results corresponds to the traditional concept of a "standard error".

If we can believe the bootstrap procedure, and we will explain later why we can, then .776 is not expected to be a highly accurate estimate of the true correlation coefficient, but not totally worthless either. Its expected level of accuracy is about one unit in the first decimal. We can't conclusively rule out that the true correlation for all law schools might be .6 or .9 , but it is certainly not zero.

In this example we can check out the answer. The true correlation for all 82 American law schools in 1973 was .761. The 15 law schools we have been discussing were randomly selected from the 82. They gave an estimate, $r = .776$, which came out close to the mark, somewhat closer in fact than would be expected on the average. The true accuracy (standard error) of r , by which we mean the variability of r when actually drawing sets of 15 at random from the 82, is ±.133 , agreeing nicely with the bootstrap estimate ±.127. Notice that only the data from the 15 selected schools entered into the bootstrap estimate, so we haven't cheated in presenting this example.

It takes about five minutes to calculate the correlation for 15 cases using a desk calculator of 1950's vintage. It takes 1/10,000 of a second to do the same calculation on a modern computer. At this speed the bootstrap becomes feasible for routine use. All of the calculations going into "±.127", drawing the 1000 bootstrap samples, computing the 1000 bootstrap correlations, and finding the



One thousand bootstrap correlation coefficients were obtained by computer resampling of the data for the original 15 law schools. The central 68% of the 1000 values includes the interval from .654 to .908. Taking half the length of this interval gives ±.127 as an accuracy measure for the observed value r = .776. The bootstrap results can be predicted theoretically, red curve, if the original data is assumed to follow a Gaussian distribution. All the bootstrap calculations together take less than a second on a medium size computer.

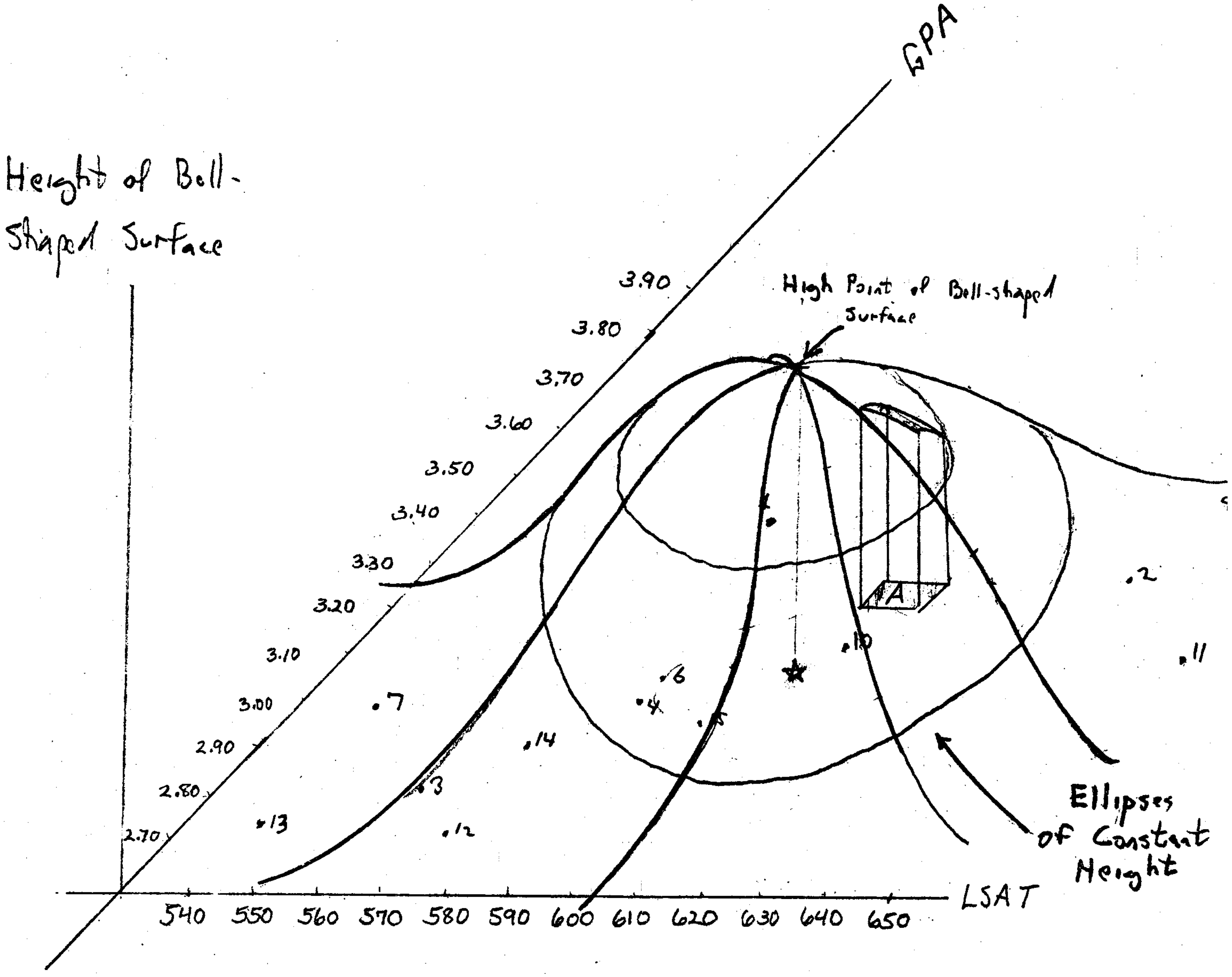
central 68% interval, takes less than a second, and costs less than a dollar. (This represents the cost of roughly 100,000 arithmetic operations. In this example we could get by, at the expense of a slightly less desirable ± estimate, with 10,000 operations. On the other hand more elaborate versions of the bootstrap, which give other kinds of information, could easily require 1,000,000 operations.)

* * *

How did statisticians calculate accuracies before computers? The answer, interestingly enough, can be couched in terms of the bootstrap. Under special circumstances the bootstrap distribution of r can be calculated theoretically. The special circumstances are that each pair of averages LSAT, GPA comes from what is known as a two-dimensional Gaussian, or "normal", or "bell-shaped" distribution. Such a distribution, the one best fitting the data for the 15 law schools is illustrated on page D.

The bootstrap algorithm pictured on page B is changed in only one way for the Guassian calculation: the slips of paper going into the hat now represent all possible points in the LSAT-GPA plane, not just the 15 points actually observed in the sample. A given point LSAT, GPA, for instance 620, 3.30 , is written on a number of slips of paper proportional to the height of the bell-shaped surface above that point.

The theoretical bootstrap distribution for the Gaussian situation, shown in red on page C, was derived in 1915 by Sir Ronald Fisher, the greatest of statisticians and originator of many of the most widely used statistical methods. The calculation is a triumphant combination of linear algebra and the Gaussian distribution. Fisher showed that the correlation coefficient was essentially the cosine of the angle between two 15 dimensional vectors formed by the 15 LSAT averages and the 15 GPA averages. Using certain symmetry properties of the multi-dimensional Gaussian distribution, and a good deal of special function theory, he was able to give an infinite series representation for the theoretical bootstrap curve. The



Two-dimensional Gaussian distribution fitted to the data for the 15 law schools: the probability that a randomly drawn pair of averages LSAT, GPA occurs in an area A of the LSAT-GPA plane is the volume of the region between A and the bell-shaped surface. The surface has its greatest height, and gives the largest probabilities, at the average of the averages "*", and has elliptical contours of equal height centered at *. The slope and direction of the ellipses are determined by the location of the 15 data points in relation to *.

actual numerical calculation of the family of possible curves, laboriously carried out on desk calculators, was not completed for several years. The awesome power of modern computers is evident on page C; the actual bootstrap results are the effective equivalent of Fisher's calculations, but applied to the actual data instead of to the theoretical Guassian distribution, and carried out by brute force in less than a second!

The Gaussian-theory estimate of accuracy for $r = .776$, half the length of the central 68% interval for the theoretical bootstrap curve, equals ±.113, agreeing reasonably well with the bootstrap estimate ±.127 and the true accuracy ±.133.

Clever and useful as it is, there are two drawbacks to the use of Fisher's result from the point of view of a statistical consumer. First of all there is no easy way to check the assumption that the 15 sample points come from a Gaussian distribution. A much larger sample, perhaps in the hundreds is needed to check this assumption. Methods like the bootstrap are called "nonparametric" because they do not make such assumptions. Parametric methods are ones which begin by assuming that the data comes from a distributional family, like the Gaussian, which can be described in terms of a small number of parameters. When such assumptions are justified, or even roughly justified, parametric methods can be much more efficient than their nonparametric counterparts but often, as in the law school example, there is no real reason to believe them.

The second drawback of classical statistical theory is more subtle, but in some ways more limiting: attention is focused on a few standard statistics like the correlation coefficient for which mathematical analysis is possible. For the most part the analyzable statistics are those based on simple linear algebra. Computer-based methods such as the bootstrap free the statistician to attack more complicated problems, using a wider variety of weapons.

In the following paragraphs we describe four progressively more elaborate applications of the bootstrap to problems in analyzing test scores, map drawing, economic forecasting, and a complex prediction problem. The reader should keep in mind that even though the statistics being bootstrapped are complicated and difficult to concisely describe, the bootstrap itself remains essentially the same simple device used in the law school example: the data in hand is used as an estimate of the population of interest, fake data sets are generated from this population, the statistic of interest is computed on the fake data set, the variability over the bootstrap replications is an estimate of the true variability.

* * *

Here is an application of the bootstrap to a problem that is right at the boundary of mathematical tractability. Eighty-eight college students each took five tests (data from Multivariate Analysis by Mardia, Kent and Bibby). It is not a simple matter to look over the list of scores - five per student - and compare the students. One approach is to replace the five numbers associated to each student by one number such as the final test score or an average of the students five scores. These numbers are both weighted averages of five scores, the final test score uses weights (0,0,0,0,1) and the average uses weights $(\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5})$. Which set of weights is most informative? Clearly, if a set of weights results in an "average" score that is essentially the same for all students, it is not useful in understanding student differences. The method of principal components, described in the figure on page G1, was introduced by the statistician Harold Hotelling in 1933. The method seeks out weighted combinations of the five test scores that vary most between the students and are therefore most useful for making comparisons. In the example, the first principal average, i.e. the single most variable linear combination, is .51 × Score 1 + .37 × Score 2 + .35 × Score 3 + .45 × Score 4 + .53 × Score 5. The weights .51, .37, .35, .45, .53 are scaled to have sum of squares 1.0. The five weights are roughly equal. Thus the principal component average can be interpreted as a student's average performance on the five tests.

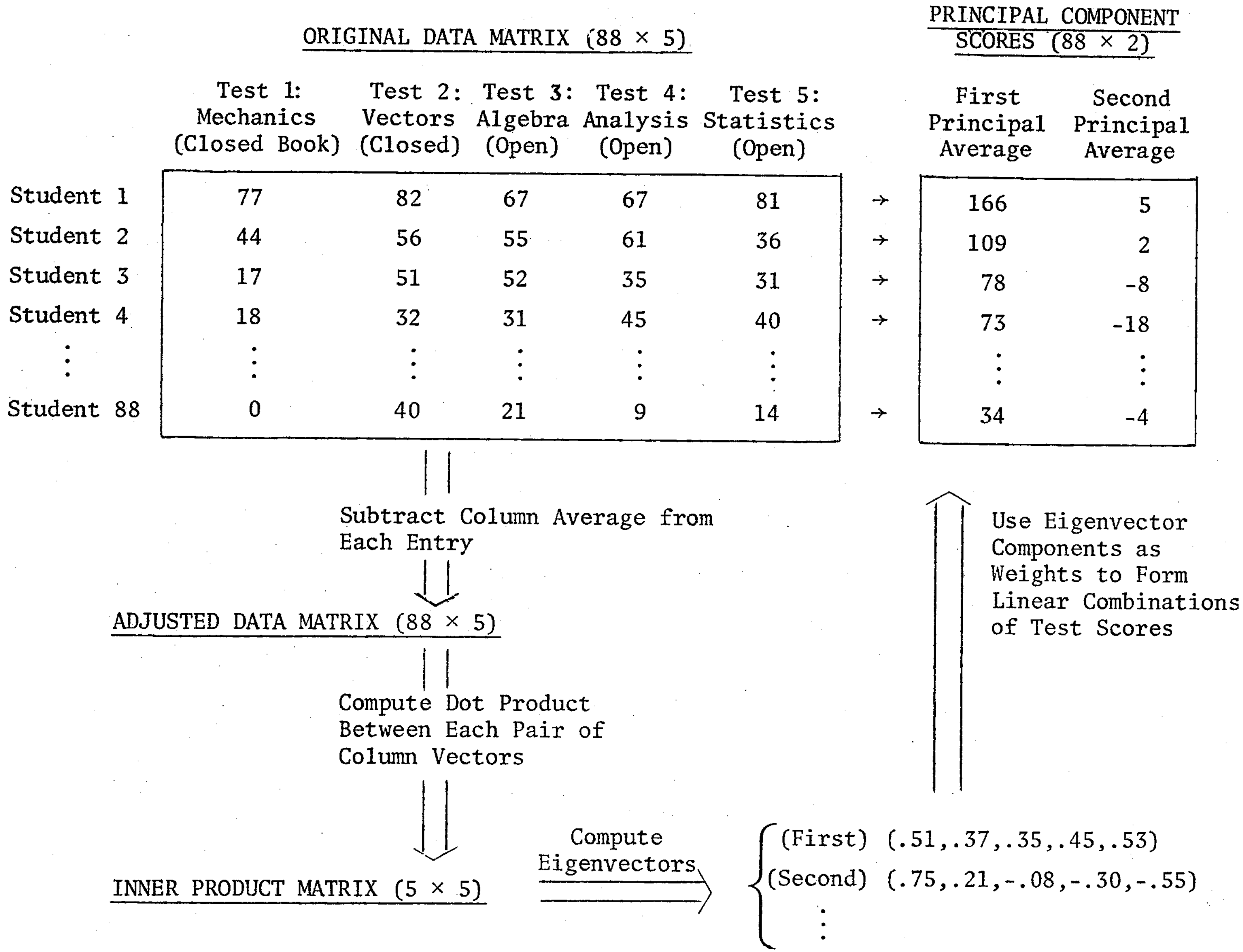
The second principal component is defined as the next most variable linear combination of the scores, subject to the constraint that its weight vector be orthogonal to that for the first principal component. In the example, it turns out to be

$$.75 \times \text{Score } 1 + .21 \times \text{Score } 2 - .08 \times \text{Score } 3 - .30 \times \text{Score } 4 - .55 \times \text{Score } 5 \ .$$

This represents a difference between a weighted average of the first two test scores and the last three test scores. The first two tests were closed book exams and the last three tests were open book exams, so the combination is easy to interpret. It measures a student's difference in performance between closed and open book exams. The two linear combinations just found might be used as a basis for giving grades for the year. Further principal components can be defined but we will focus on just the first two.

Given that the principal components suggest interesting interpretations of the data, we are led to the basic problem of statistical inference: If the data had come from a different 88 students, would the final inference have changed dramatically? The problem of quantifying the sampling variability of principal components has occupied many mathematical statisticians during the past 50 years. One approach is to consider small perturbations of the data by means of an appropriate bell-shaped curve. The problem then becomes a completely specified mathematical question. Today there are only partial solutions to questions concerning the sampling distribution of the first principal component. The answers involve beautiful mathematics connected to integrals over the group of orthogonal matrices. Little is known about the second and higher components. At this point one might well wonder whether the second principal component found for the 88 students has genuine meaning, or is just an artifact of sampling variability.

It is straightforward to answer this question using the bootstrap: Each student's five scores are repeatedly copied onto many pieces of paper. A new sample



ORIGINAL DATA MATRIX (88 × 5)
PRINCIPAL COMPONENT SCORES (88 × 2)
Subtract Column Average from Each Entry
ADJUSTED DATA MATRIX (88 × 5)
Compute Dot Product Between Each Pair of Column Vectors
INNER PRODUCT MATRIX (5 × 5)
Compute Eigenvectors
(First) (.51,.37,.35,.45,.53)
(Second) (.75,.21,-.08,-.30,-.55)
Use Eigenvector Components as Weights to Form Linear Combinations of Test Scores

ORIGINAL DATA MATRIX (88 × 5)

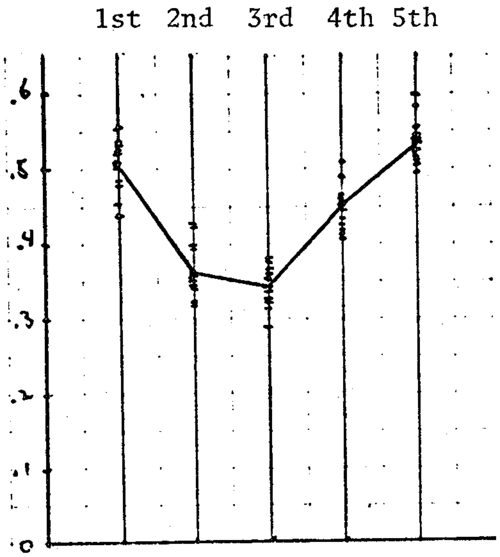
	Test 1: Mechanics (Closed Book)	Test 2: Vectors (Closed)	Test 3: Algebra (Open)	Test 4: Analysis (Open)	Test 5: Statistics (Open)
Student 1	77	82	67	67	81
Student 2	44	56	55	61	36
Student 3	17	51	52	35	31
Student 4	18	32	31	45	40
⋮	⋮	⋮	⋮	⋮	⋮
Student 88	0	40	21	9	14

PRINCIPAL COMPONENT SCORES (88 × 2)

First Principal Average	Second Principal Average
166	5
109	2
78	-8
73	-18
⋮	⋮
34	-4

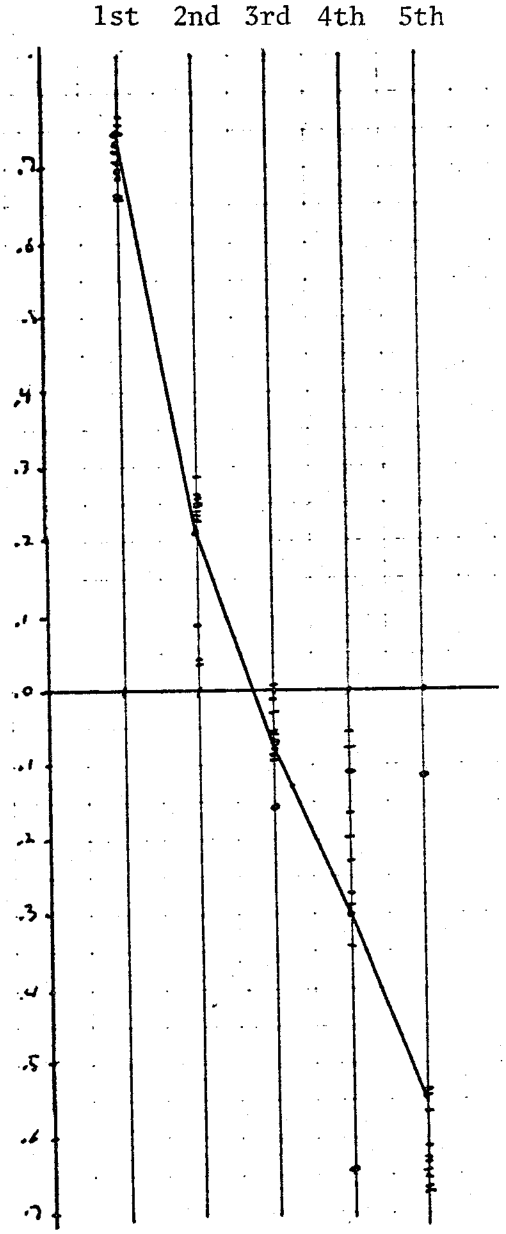
88 students each took 5 tests. Principal components is a method for finding the most informative linear combinations of the test scores. The method is based on algebraic manipulations of the original data matrix. The computations take several hours on a desk calculator, and about 1/100 of a second on a computer. What is the sampling variability of the computed eigenvectors? The figure on page G2 shows the bootstrap answer to this question. (The student test data appears in full in Multivariate Analysis by Mardia, Kent, and Bibby.)

First Eigenvector Components



Values of bootstrap eigenvalue components indicated by dashes.

Second Eigenvector Components



The bootstrap was used to assess the variability of the eigenvector components (principal component weights) found for the student-test data. The bootstrap calculations are the same as the original calculations described on page G1, except that the original data matrix is replaced by a bootstrap data matrix. This is an 88 × 5 matrix obtained by writing each student's data on many slips of paper as on page B, and drawing a bootstrap sample of size 88 from the hat. Following through the steps on page G1 gives the bootstrap eigenvectors. Repeating this whole process many times gives many sets of bootstrap eigenvectors and the variation of these is a dependable measure of accuracy for the actual eigenvectors. The figure shows the results of the first 10 bootstrap replications. The values of the bootstrap eigenvector components are indicated by dashes. (The red lines trace the actual eigenvector components.) Notice the instability in the 4th and 5th components of the second eigenvector.

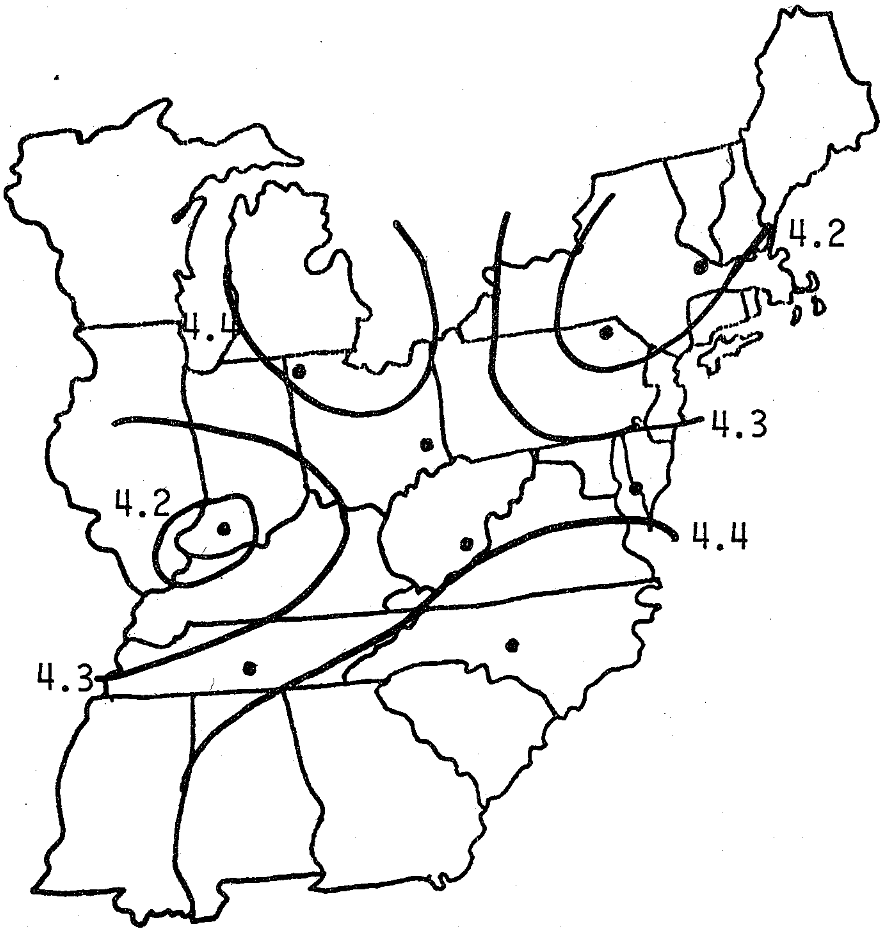
of size 88 is drawn, and the principal components are calculated. This bootstrap resampling is repeated many times. The variability of the bootstrap principal components reflects the sampling variability of the actual principal components. The results, shown on page G2, suggest the following: the weights associated with the first principal component are quite stable - they vary in their second decimal place. The weights associated with the second principal component are less stable, but in a structured way. Recall that the second principal component represented a difference between an average of the open and closed book tests. This interpretation stands up in the bootstrap analysis, but the weights within each of the two types of tests bounce around a fair amount, and shouldn't be taken too seriously.

The bootstrap analysis is easy to do, and takes about two seconds of computer time for 100 bootstrap replications. It is valid without assuming a Gaussian distribution for the population of student scores. It shows the computer together with a simple idea rolling right over a problem that mathematical theory finds almost intractable.

* * *

Not every statistic is a number. The map on page H, prepared by Barry Eynon and Paul Switzer of the Stanford statistics department, relates to acid rainfall in the Eastern United States. Nine weather stations recorded the pH level (low pH indicating high acidity) of every rainfall from September 1978 through August 1980. The contours of equal pH level show particularly heavy acidity around Scranton, Pennsylvania and around Rockport, Indiana, with a corridor of less acidity in between. How accurate are those contours? A bootstrap analysis gives interesting answers.

The contours on page H were calculated by a data-fitting procedure called "Kriging" (after South African mining engineering H. G. Krige), a computer-based map-drawing algorithm. In broad terms, the procedure (1) Begins with the 2000 pH values recorded at the nine stations during the two years, (2) Subtracts the



Nine weather stations (red dots) in Eastern United States recorded the pH level of every rainfall from September 1978 to August 1980. Low pH level indicates high acidity. Contours of equal pH level were calculated by Kriging, a computer-based fitting technique which brings all the data to bear on the estimated pH level at any one location. The contours of greatest acidity, pH = 4.2 , are near the stations at Scranton, Pennsylvania (upper right) and Rockport, Indiana (middle left). The center of the map shows a four-branched corridor of lower acidity. This map, and the bootstrap analysis shown on page I, were prepared by Barry Eynon and Paul Switzer of Stanford University.

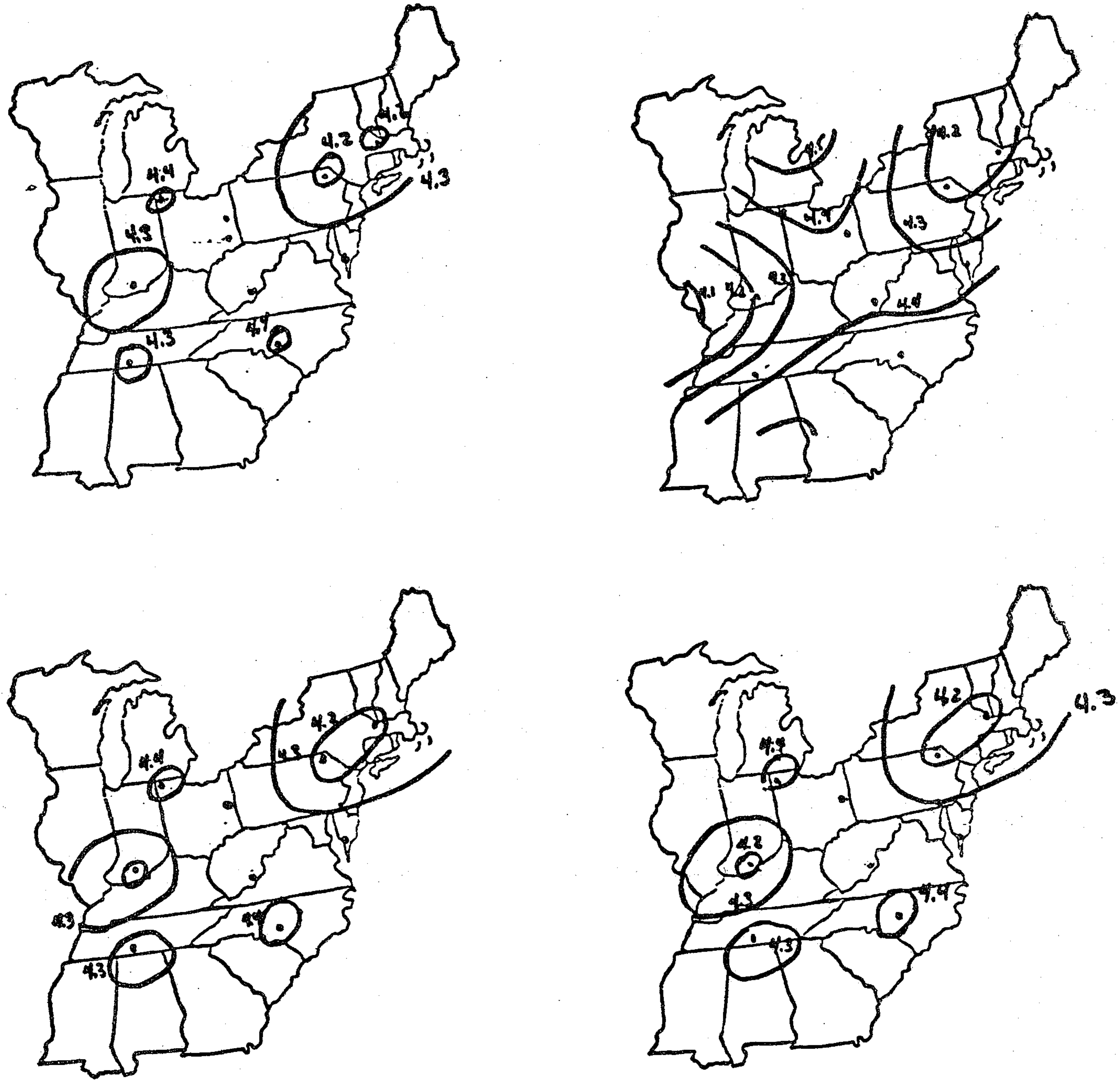
seasonal variability at each station, leaving 2000 "residuals" whose values reflect geographical differences in pH levels, (3) Fits a simple model of spatial correlation between the residuals at different points. This model describes how a pH level observed at point x on the map correlates with a simultaneously observed pH level say 50 miles north and 30 miles east of x. (4) Uses the spatial correlation function, and the observed average pH levels at the nine stations, to draw the contours. (This last step is the actual Kriging. If the spatial correlation function is known, rather than estimated, it is the optimal way to draw the contours.)

How accurate is the map on page H? Would another similar but independent set of 2000 readings give nearly the same estimators or something wildly different? The bootstrap answer is shown on page I. By resampling the 2000 residual pH values, and then following through steps (3)-(4) above on the resampled data, Eynon and Switzer constructed the bootstrap maps.

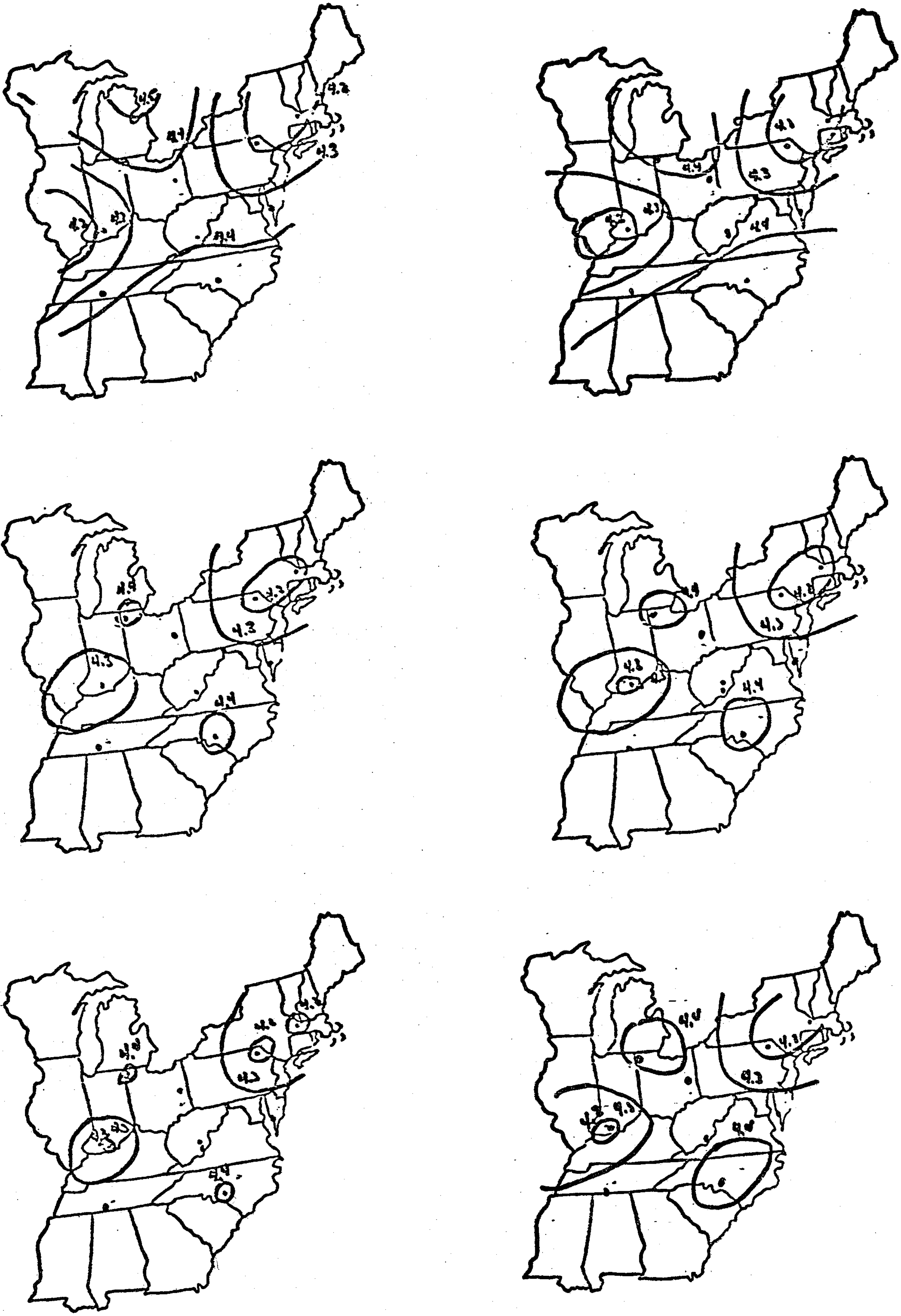
The variability among the ten bootstrap maps is striking, and shows that the original map must be interpreted cautiously: even though it is based on an enormous data set, and constructed according to a near-optimal map-drawing algorithm, it still suffers from considerable statistical variability. Instead of a corridor of low acidity we might quite possibly have seen isolated islands of high acidity, depending on the whims of random noise.

* * *

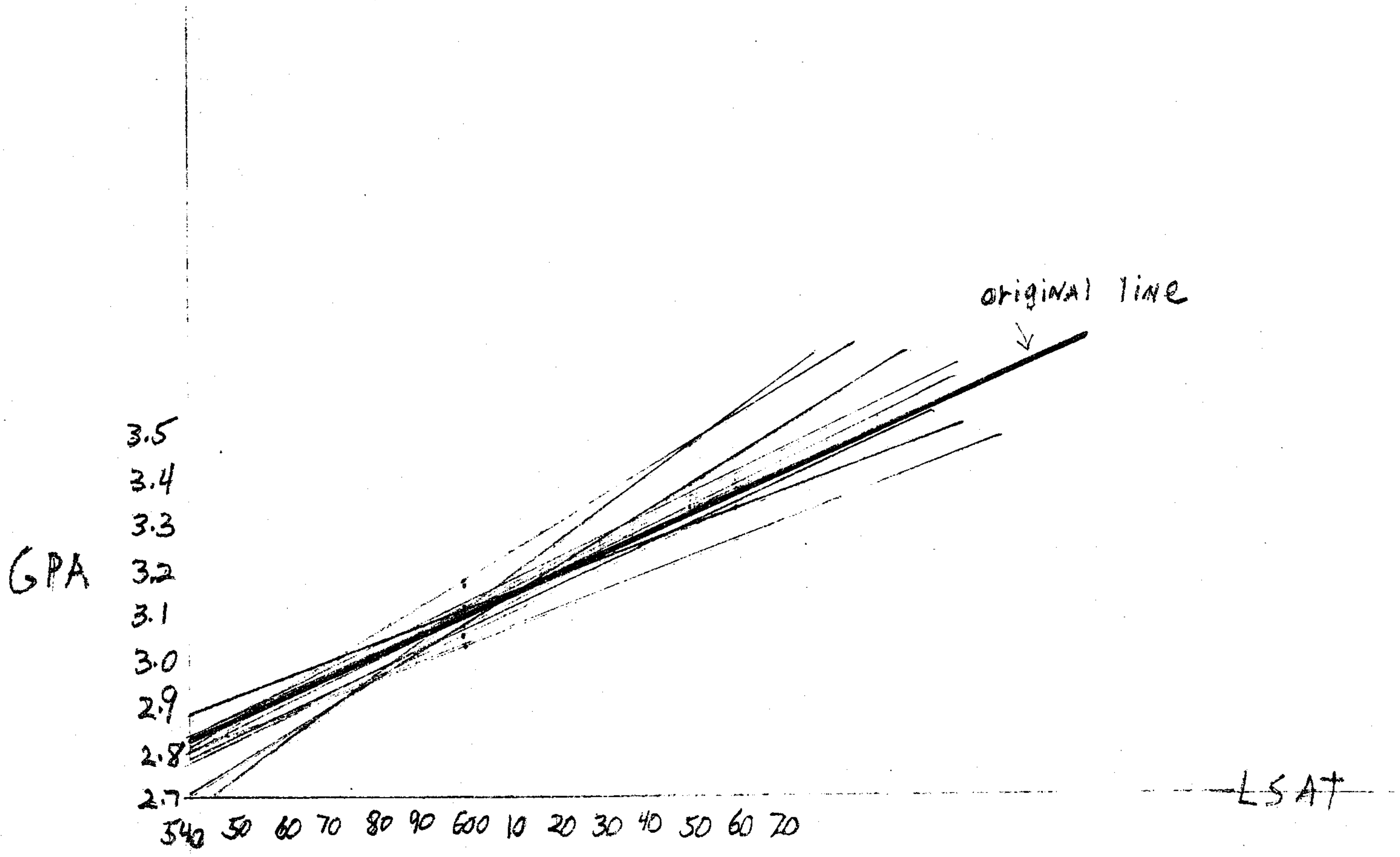
The next example shows how to use the bootstrap to understand the variability in the most widely used curve fitting algorithm. Least squares is an objective form of curve-fitting. Objectivity is enforced by agreeing before the fact exactly how the curve is going to fit to the data. The figure on page J shows a simple example relating to the law school data of page A. Here the curve being fit is a straight line. The "original line" in figure J is the least squares line. By definition this is the straight line minimizing the sum of squared vertical discrepancies between the line and the 15 data points.



How much should we trust the pH contours shown on the map on page H? Ten bootstrap maps are shown here. They were obtained by resampling the 2000 residual pH values (after subtracting the seasonal variability at each station) and applying the Kriging procedure to the resampled data. Some of the bootstrap maps look much different than the original; instead of a corridor of low acidity they show isolated islands of high acidity. The original map must be interpreted with caution. Even though it is based on a huge amount of data, its contours are still subject to considerable statistical variability.


4.5
4.4
4.2
4.3
4.2
4.2
4.4
4.1
4.4
4.3
4.2
4.1
4.4
4.5
4.2
4.3
4.5
4.4
4.4
4.2
4.3
4.2
4.4
4.2
4.2
4.4
4.4
4.3
4.3
4.2
4.3
4.4

I (cont'd)



The "original line" is the linear regression of GPA on LSAT for the 15 law schools. It is the line of form a + b • LSAT which minimizes the sum of squares of the vertical discrepancies GPA - (a+b•LSAT), summed over the 15 data points. The ten bootstrap replications of the least squares fitting process indicate that the original line is subject to considerable statistical variability, but not so much as to make it useless for predicting GPA from LSAT.

Gauss and Lagrange invented least squares in the early 1800's, for use in astronomical predictions. In the law school example, the original line might be used to predict GPA from LSAT, if the latter but not the former could be observed. The ten bootstrap lines show that these predictions would contain a large component of statistical variability, but not so much as to make them useless. (In this case, which is particularly simple, the bootstrap can be analyzed theoretically, and shown to give almost the same results as Gauss Lagrange's original mathematical treatment.)

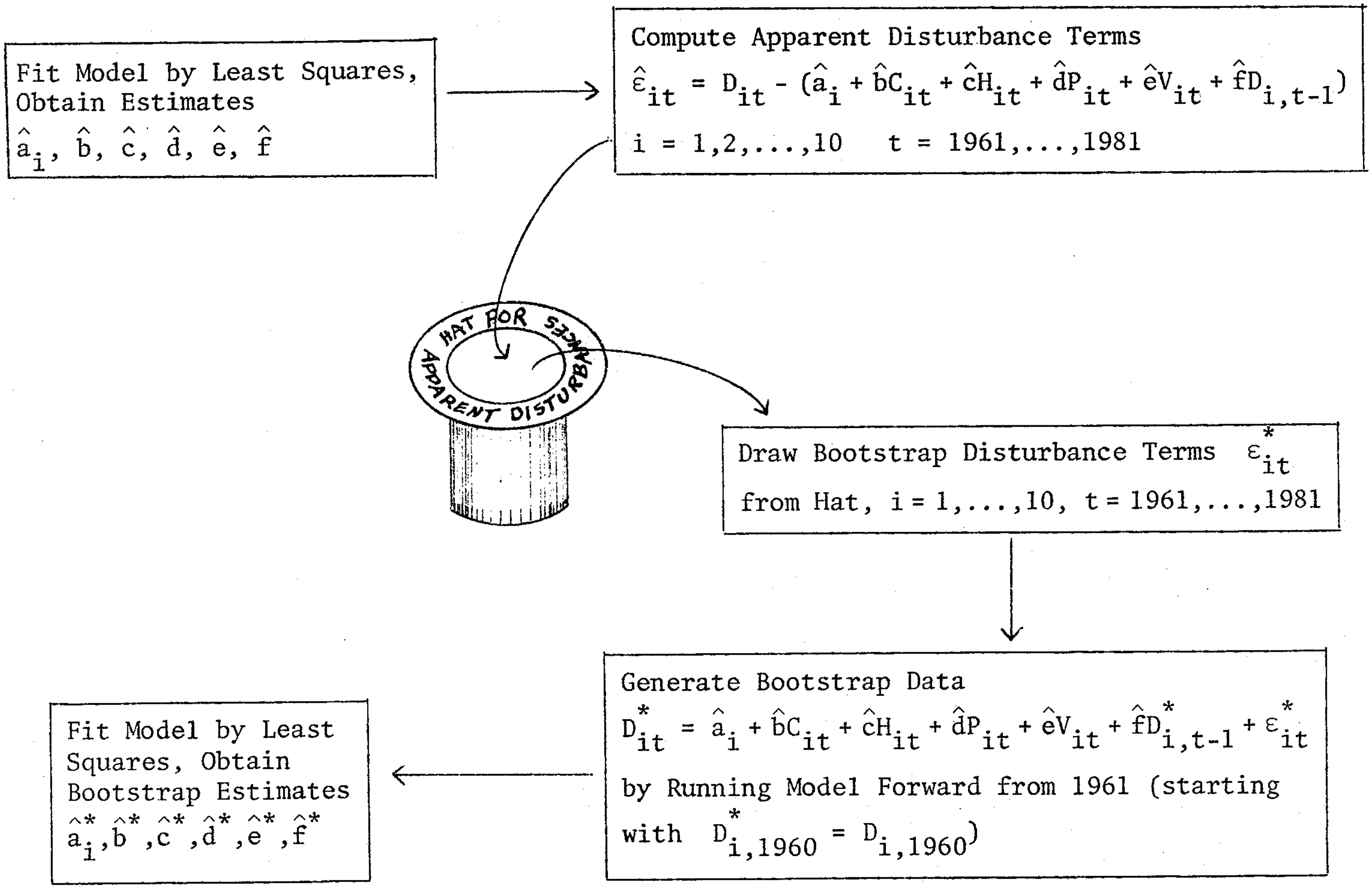
Least squares may be the most widely used statistical method. It is particularly useful in complicated situations where the scientist needs to bring large amounts of diverse information to bear on a single question. As an important example, consider the Department of Energy's forecasts of energy demand. In one version, called RDFOR, energy demand is modelled by a curve fit to the previous year's demand, and related measured quantities such as weather and fuel price.

Part of the RDFOR model is described in the figure on page K. Of course the model is not expected to fit the data perfectly. Allowance for this is made by including "stochastic disturbance terms", which is another name for errors. The model is fitted by choosing values for the unknown parameters which minimize, in a generalized sense, the sum of squares of the apparent disturbances. The fitted values for the unknown parameters are the best available estimates for these important quantities. How accurate are they?

David Freedman and Stephen Peters, of the Berkeley and Stanford statistics departments respectively, answered the question by the bootstrap analysis indicated on page K. This first involved creating bootstrap data sets by resampling the apparent disturbances and adding them back to the originally fitted RDFOR model. Bootstrap values of the parameter estimates were obtained by applying the least

Bootstrapping an Energy Model

Model: $D_{it} = a_i + bC_{it} + cH_{it} + dP_{it} + eV_{it} + fD_{i,t-1} + \varepsilon_{it}$



Fit Model by Least Squares, Obtain Estimates $\hat{a}_i, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{f}$

Compute Apparent Disturbance Terms
$\hat{\varepsilon}_{it} = D_{it} - (\hat{a}_i + \hat{b}C_{it} + \hat{c}H_{it} + \hat{d}P_{it} + \hat{e}V_{it} + \hat{f}D_{i,t-1})$
$i = 1,2,\ldots,10 \quad t = 1961,\ldots,1981$

Draw Bootstrap Disturbance Terms ε^*_{it} from Hat, $i = 1,\ldots,10$, $t = 1961,\ldots,1981$

Generate Bootstrap Data
$D^*_{it} = \hat{a}_i + \hat{b}C_{it} + \hat{c}H_{it} + \hat{d}P_{it} + \hat{e}V_{it} + \hat{f}D^*_{i,t-1} + \varepsilon^*_{it}$
by Running Model Forward from 1961 (starting with $D^*_{i,1960} = D_{i,1960}$)

Fit Model by Least Squares, Obtain Bootstrap Estimates $\hat{a}^*_i, \hat{b}^*, \hat{c}^*, \hat{d}^*, \hat{e}^*, \hat{f}^*$

The model RDFOR is used by the Department of Energy to analyze and forecast energy demand. The country is divided into 10 regions. The primary data fitted by the model are D_{it}, logarithm of energy demand in region i for year t, $i = 1,2,\ldots,10$, $t = 1961,\ldots,1981$. A regression equation, shown at the top of the figure links D_{it} to the previous year's demand $D_{i,t-1}$, and also to various measured variables: C_{it} is log degree cooling days, H_{it} is log degree heating days, P_{it} is log fuel prices, and V_{it} is log value added in manufacturing. The "stochastic disturbance terms" ε_{it} are unobservable prediction errors. The unknown parameters a_i, b, c, d, e, f are estimated by generalized least squares fitting of the model to the data, giving estimates $\hat{a}_i, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{f}$. How accurate are these estimates? Bootstrap analysis by David Freedman, Berkeley, and Stephen Peters, Stanford, showed that they are two or three times more variable than indicated by standard approximation theory.

squares fitting procedure to the bootstrap data. The variability of the bootstrap values gave a direct estimate of the accuracy of the original estimates.

The usual estimates of accuracy for this situation are based on a mathematical approximation theory called generalized least squares. The bootstrap estimates of variability (standard error) showed that these approximations were much too small, by factors of two or three in most cases. Accuracy of the parameter estimates is a crucial factor in how well RDFOR predicts future energy demand, the main reason for its existence. Knowing that these estimates are substantially less dependable than previously thought is disappointing but important information.

* * *

The examples presented so far have involved clearly posed problems. In practice, statistical procedures are not always completely specified before the data is examined. Indeed, several analyses may be performed upon the same data set. The next example illustrates the bootstrap in action on a real problem with these features. The data was provided by Dr. Peter Gregory of Stanford University's School of Medicine. The analysis is based on the Ph.D. dissertation of Gail Gong at Stanford.

The scientific problem involve 155 acute chronic hepatitis patients. Of these, 33 were observed to die from the disease while 122 survived. For each patient, 19 measured variables were available. These are potentially useful predictors like age, sex, and standard chemical measurements. Data from the last 11 patients is shown in the figure on page G. Dr. Gregory's aim was to understand the effect of the measured variables on the chance of patient survival. The analysis involved several steps: statistical experience suggests that it is unwise to fit a curve depending on all of the 19 measured variables with only 155 patient records available. Thus the first job was to reduce the data to 4 or 5 important variables. This itself was done in two stages: first, by preliminary inspection of each variable separately, and then by an automated procedure known as stepwise

patient #	age	sex	steroid	antiviral	fatigue	malaise	anorexia	liver big	liver firm	spleen palp	spiders	ascites	varices	bilirubin	alk phos	SGOT	albumin	protein	prognosis	outcome
145	45	M	2	2	1	1	1	2	2	2	1	1	2	1.90	*	114	2.4	*	*	die
146	31	M	1	2	1	2	2	2	2	2	2	2	2	1.20	75	193	4.2	54	2	survive
147	41	M	2	2	1	2	2	2	1	1	1	2	1	4.20	65	120	3.4	*	*	die
148	70	M	1	2	1	1	1	*	*	*	*	*	*	1.70	109	528	2.8	35	2	die
149	20	M	1	2	2	2	2	2	*	2	2	2	2	0.90	89	152	4.0	*	2	survive
150	36	M	2	2	2	2	2	2	2	2	2	2	2	0.60	120	30	4.0	*	2	survive
151	46	M	2	2	1	1	1	2	2	2	1	1	1	7.60	*	242	3.3	50	*	die
152	44	M	2	2	1	2	2	2	1	2	2	2	2	0.90	126	142	4.3	*	2	survive
153	61	M	1	2	1	1	2	1	1	2	1	2	2	0.80	95	20	4.1	*	2	survive
154	53	F	1	2	1	2	2	2	2	1	1	2	1	1.50	84	19	4.1	48	*	survive
155	43	M	2	2	1	2	2	2	2	1	1	1	2	1.20	100	19	3.1	42	2	die

155 chronic hepatitis patients were observed to die from the disease or to survive. 19 Predictor variables were recorded for each patient, though some of the data, indicated by "*", was missing. Data for the last 11 patients is shown above. Dr. Peter Gregory of Stanford used this data to fit a rule for predicting survival. The fitted rule chose malaise, ascites, bilirubin, and prognosis as the important predictor variables, and correctly predicted 84% of the 155 patients. His entire data analysis procedure, including preliminary steps, was subject to bootstrap analysis to assess the variability of the conclusions. See figure on page M.

multiple logistic regression. Four variables survived these steps, those labelled malaise, ascites, bilirubin, and prognosis. The next job involved fitting a curve that predicts the proportion of surviving patients as a function of these four variables. This curve, obtained by a close cousin of least squares fitting, is the actual logistic regression. The logistic regression curve gives a quantitative prediction of chance of survival as a certain mathematical function of the patient's measured malaise, ascites, bilirubin, and prognosis. The analysis implies that these variables are the main determining factors and this in turn suggests scientific consequences.

The analysis described above is a complex, multistaged procedure. It is typical of scientific practice. How sure can we be of the conclusions drawn from this analysis? After all, there is random error in the variables and the choice of the study population, upon which we have superimposed an elaborate fitting technique. The bootstrap offers some indication of the variability of this whole complex procedure. The idea is to bootstrap the entire data analysis, starting from the very beginning. Thus a sample of 155 records is drawn with replacement from the original set of 155 records. Then, these records are analyzed by the same sequence of procedures outlined above; starting with a preliminary screening for important variables and ending with a logistic curve fitted through the final variables. The curve and variables are recorded and a new sample of 155 is drawn.

The results are surprising and informative. The final set of "important variables" is iteslf quite variable. For example, no single variable appeared in 50 percent of the bootstrap replications. The figure on page M contains further details.

Bootstrapping lends insight into other aspects of this example. The original analysis resulted in a fitted curve that predicts the chance of death as a function of measured variables. How accurate is this prediction rule? A naive answer to

Bootstrap #	Variables Selected
476	ascites, malaise, histology, bilirubin
477	ascites, protein, fatigue
478	histology, alk phos, protein
479	histology, protein
480	varices, albumin, malaise, alk phos, age
481	albumin, histology, malaise, spleen palp
482	histology, protein, bilirubin
483	histology
484	ascites, spiders, bilirubin, anorexia, albumin, malaise, protein
485	bilirubin, ascites, protein
486	ascites, steroid
487	spiders, bilirubin, sex
488	bilirubin, alk phos, sex
489	bilirubin, histology, steroid
490	alk phos, ascites, age, protein
491	albumin, histology, sex
492	ascites, bilirubin, histology
493	bilirubin, ascites
494	bilirubin, histology, malaise
495	ascites
496	bilirubin
497	ascites, varices
498	spiders, histology, albumin
499	age, histology, bilirubin, malaise, protein, spiders
500	ascites, histology, bilirubin, protein

Variables selected as "important" in the last stage of the hepatitis example. This table gives the results of the last 25 bootstrap replications. The variables selected by the original data were malaise, ascites, bilirubin, and prognosis. In 500 bootstrap replications, malaise was selected 35% of the time. The other proportions are ascites - 37%, bilirubin - 48%, prognosis - 59% of the time. No theory exists for interpretation of these results but they discourage taking the final variables in the original analysis too seriously. Bootstrap analysis by Dr. Gail Gong, Carnegie-Mellon University, with the help of Dr. Gregory.

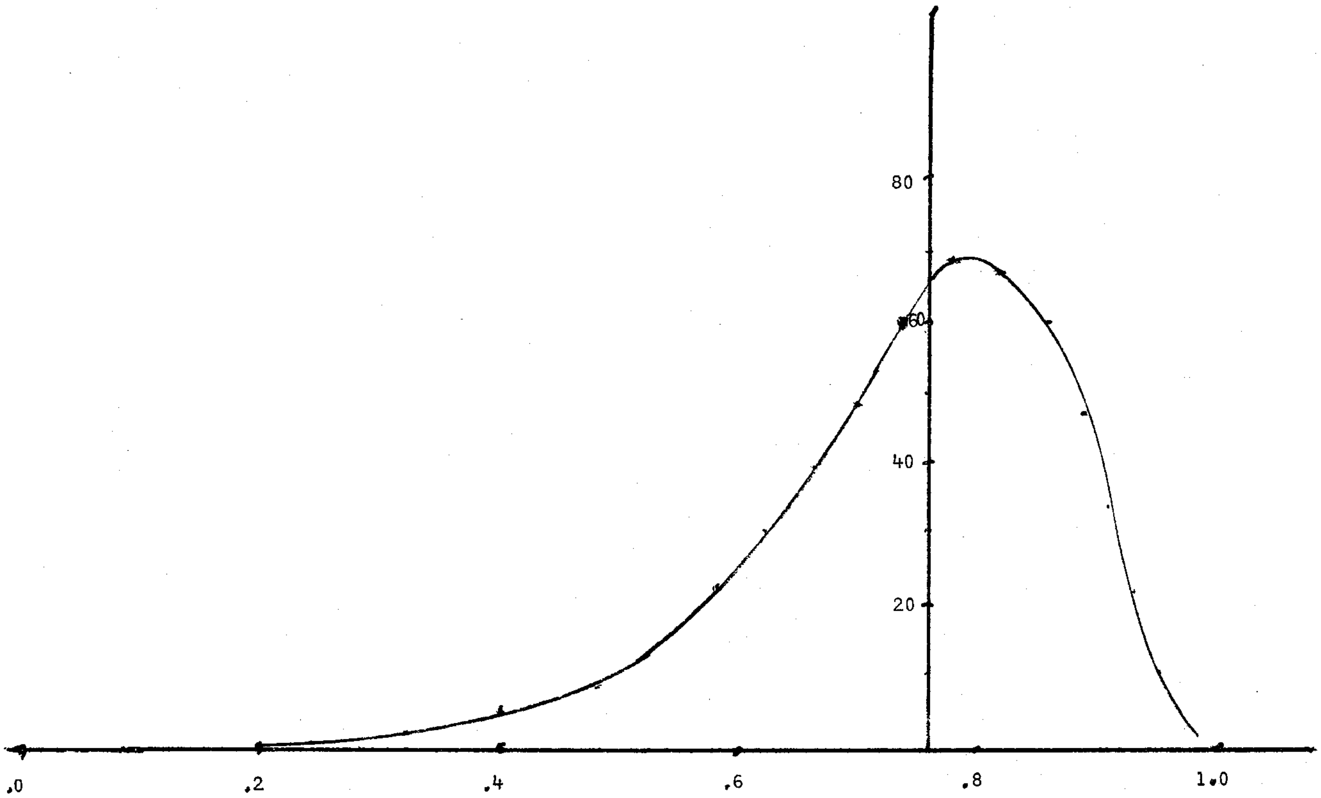
this question is to try the rule out on the data that generated it: look at each of the 155 original patients and predict survival if the fitted curve gives chance 50% or more to survival. This rule misclassified 16% of the 155 patients. However 16% underestimates the true probability of misclassification because the same data is used both to generate and evaluate the rule. The bootstrap analysis can correct for this effect, and suggests that a better estimate for the misclassification rate is 20%.

The prospect of bootstrapping at the level of entire data analyses offers hope in a very hard problem - the connection between the mathematical theory underlying statistics and actual statistical practice. Much practical work with data begins by graphing data, removing obvious errors, and making preliminary inspections of the situation. The effect of this kind of "data-snooping" is usually ignored in statistical analyses, for no better reason than that it is impossible to mathematically analyze. The bootstrap using the power of the computer, assesses the effects of these preliminary operations on the final conclusion.

* * *

In the discussion so far, the bootstrap has been motivated as a common sense method of getting a handle on variability. We will now present a more careful account of what it means to say the bootstrap "works". The basic idea is simple - the bootstrap has been tried out in a large number of problems where the correct answer is known. The bootstrap works in these problems and can be mathematically proven to work in similar problems. The word "mathematically" is important here. Computer-based methods like the bootstrap don't eliminate mathematical thinking, they simply change the level at which mathematics is applied. Mathematical statisticians have been busy investigating how well we can expect methods like the bootstrap to work in general.

Consider the law schools again. There are 82^{15} equally likely ways of choosing a random sample of size 15 from the 82 schools, each of which gives a



Actual Distribution of r , 15 law schools drawn with replacement from the 82.

The actual distribution of the correlation coefficient r for the law schools, based on drawing 1,000,000 random samples of size 15 from the 82 law schools. Except for being smoother, because of the increased number of correlations plotted, the curve looks much like the bootstrap results shown in the figure on page C. This is no accident. Theoretical calculations carried out by R. Beran, P. Bickel, and D. Freedman (Berkeley), and by K. Singh (Rutgers) show that the bootstrap and actual results tend to agree with each other to a high order of approximation.

value of the correlation coefficient r. The actual distribution of these values is shown in the figure on page N. (This figure is based on 1,000,000 randomly selected samples of size 15, which is enough to eliminate any discernible differences from the picture based on all 82^{15}.)

The bootstrap distribution shown in the figure on page C looks much like the actual distribution shown on page N, and would look even more like it if we had taken 1,000,000 rather than 1,000 bootstrap samples. This is no accident. Theoretical work carried out by Rudolf Beran (Berkeley), Peter Bickel and David Freedman (Berkeley), and Kesar Singh (Rutgers) show that,properly centered, the two curves will tend to agree to a surprisingly high order of approximation.

In particular the bootstrap reveals the correct asymmetry of the sampling distribution - in this case the long tail to the left of the central value evident on both pages C and N. With this information we can do better than assign a simple ± accuracy estimate. We can construct what are known as "confidence intervals": e.g. based on the bootstrap data in the figure on page C, the true correlation coefficient for the entire population of law schools has probability about 68% of lying in the interval (.606, .876).

Current theoretical work focuses on three aspects of the bootstrap: to justify the agreement between actual and bootstrap sampling distributions under the widest possible circumstances, to use this agreement to obtain valid confidence statements, and finally to investigate possible improvements on the bootstrap.

It is worth noticing that the theoretical bootstrap curve on page C, derived under the Gaussian assumption, is also a good approximation to the actual distribution of r shown on page N. Gaussian theory is a notable over-acheiver, and as in the law school case usually gives good results even if the Gaussian assumptions are only satisfied in the crudest sense. That is why applied statisticians have gotten by quite well without computers.

The real objection to Gaussian theory, and the real advantage of computational methods like the bootstrap, is the need for theory, not the Gaussian assumptions. It is hard work deriving results like Fisher's theoretical distribution on page C, even for very simple statistics, and impossible work for the more complicated statistics in our examples. The computer is not subject to the limitations of mathematical simplicity.

* * *

We have mentioned the interest in possible improvements on the bootstrap. There exist a number of alternative computer-based methods which are similar in spirit but quite different in detail. Those in current use include the jackknife, cross-validation, and balanced repeated replications. Each of these methods generate fake data sets from the original data, and use the variability of the results from the fake data to assess the actual variability in the original results. They differ from the bootstrap and from each other in how the fake data sets are generated.

The jackknife was the first such method, created in the 1950's by Maurice Quenouille (deceased) and John Tukey (Princeton and Bell Labs), and extensively investigated by Colin Mallows (Bell Labs), Louis Jaeckel (Berkeley), David Hinkley (Austin), Rupert Miller (Stanford), William Schucany (Texas A & M), and many others. The name jackknife was coined by Tukey to suggest a useful all-purpose statistical tool.

The jackknife proceeds by removing one observation at a time from the original data, recalculating the statistic of interest, and seeing how it varies as the removed observation goes through all possibilities. For the law school data it gives a ± estimate based on 15 recalculations of r. The jackknife is less of a computational spendthrift than the bootstrap, but also seems less flexible and, sometimes, less dependable.

Cross-validation is an elaboration of the following simple idea. Set aside half of the data. Fit curves to the first half to your heart's content, choose the one that fits best, and then try this curve out on the second half. This last step, the cross-validation, gives a dependable indication of how the fitted curve would perform on new data. There is nothing special about half splits - 90% and 10% can do as well. And there is no reason to cross-validate only once. The data can be randomly split in half many times.

Cross-validation has been widely applied in situations where a curve-fitting procedure is well-defined by standard theory except for a crucial parameter, usually relating to the smoothness of the curve. For instance we may be willing to fit a polynomial to the data by least squares, but not know which degree polynomial to fit. Cross-validation will choose that degree polynomial, fit to the first half of the data, which performs best on the second half. Seymour Geisser (Minnesota), Mervyn Stone (London) and Grace Wahba (Wisconsin) have been pioneers in this development.

Instead of splitting data into halves at random, a more systematic system of splits can be used. These splits can be carefully chosen so that the results are optimal in certain simple situations which permit full theoretical analysis. This is the idea underlying the balanced repeated replication method of Philip J. Mccarthy at Cornell. This method is widely used in assessing variability in surveys and census samples. Random subsampling, a related method developed by John Hartigan (Yale), is specifically designed to yield dependable confidence intervals in some situations.

There are close interconnections linking all of these methods. One line of development derives them all, and others not mentioned here, from the bootstrap. Over the past few years statisticians have been working out these connections, seeking improvements, and connecting the results to classical Gauss-Fisher theory.

We return to our original point: the computer is changing statistical theory. Above we have focused on new theories that have evolved because of the computer. Another obvious change is the huge data bases that become available because of computer memory. Also, the computer enables us to use the old methods to solve larger problems. Principal components is a nice example of this; the method was invented before the computer but not really usable.

It must also be said that there are some problems that even the biggest computers cannot solve by brute force. Examples include modern approaches to cryptography and the following simple problem - how many times do we have to shuffle a deck of cards until it is close to random? Here the point is that the number of possible arrangements of a deck of cards, 52! , or about 10^{67} , is just too large for a computer to handle. On the other hand the computer coupled with theory do just fine on this problem - about 7 ordinary shuffles are enough.

Fisher was able to provide a statistical theory which took full advantage of the computational facilities of the 1930's. The goal now is to do the same for the 1980's.

References

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. SIAM Review **21**, 460-480.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* *SIAM* Monograph #38, NSF-CBMS.

Eynon, B., and Switzer, P. (1982). The variability of acid rainfall. Technical Report No. 58, Department of Statistics, Stanford University.

Freedman, D., and Peters, S. (1982). Bootstrapping a regression equation: some empirical results. Technical Report No. 10, Department of Statistics, University of California, Berkeley.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.