Stein's Paradox in Statistics

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

Sometimes a mathematical result is strikingly contrary to generally held belief even though an obviously valid proof is given. Charles Stein of Stanford University discovered such a paradox in statistics in 1955. His result undermined a century and a half of work on estimation theory, going back to Karl Friedrich Gauss and Adrien Marie Legendre. After a long period of resistance to Stein's ideas, punctuated by frequent and sometimes angry debate, the sense of paradox has diminished and Stein's ideas are being incorporated into applied and theoretical statistics.

Stein's paradox concerns the use of observed averages to estimate unobservable quantities. Averaging is the second most basic process in statistics, the first being the simple act of counting. A baseball player who gets seven hits in 20 official times at bat is said to have a batting average of .350. In computing this statistic we are forming an estimate of the player's true batting ability in terms of his observed average rate of success. Asked how well the player will do in his next 100 times at bat, we would probably predict 35 more hits. In traditional statistical theory it can be proved that no other estimation rule is uniformly better than the observed average.

The paradoxical element in Stein's result is that it sometimes contradicts this elementary law of statistical theory. If we have three or more baseball players, and if we are interested in predicting future batting averages for each of them, then there is a procedure that is better than simply extrapolating from the three separate averages. Here "better" has a strong meaning. The statistician who employs Stein's method can expect to predict the future averages more accurately no matter what the true batting abilities of the players may be.

Baseball is a sport with a large and carefully compiled body of statistics, which supplies convenient material for illustrating the workings of Stein's method. As our primary data we shall consider the batting averages of 18 major-league players as they were recorded after their first 45 times at bat in the 1970 season. These were all the players who happened to have batted exactly 45 times the day the data were tabulated. A batting average is defined, of course, simply as the number of hits divided by the number of times at bat; it is always a number between 0 and 1. We shall denote each such average by the letter y.

The first step in applying Stein's method is to determine the average of the averages. Obviously this grand average, which we give the symbol $\bar{y}$, must also lie between 0 and 1. The essential process in Stein's method is the "shrinking" of all the individual averages toward this grand average. If a player's hitting record is better than the grand average, then it must be reduced; if he is not hitting as well as the grand average, then his hitting record must be increased. The resulting shrunken value for each player we designate z. This value is the James-Stein estimator of that player's batting ability, named for Stein and W. James, who together proposed a particularly simple version of the method in 1961. Stein's paradox is simply that the z values, the James-Stein estimators, give better estimates of true batting ability than the individual batting averages.

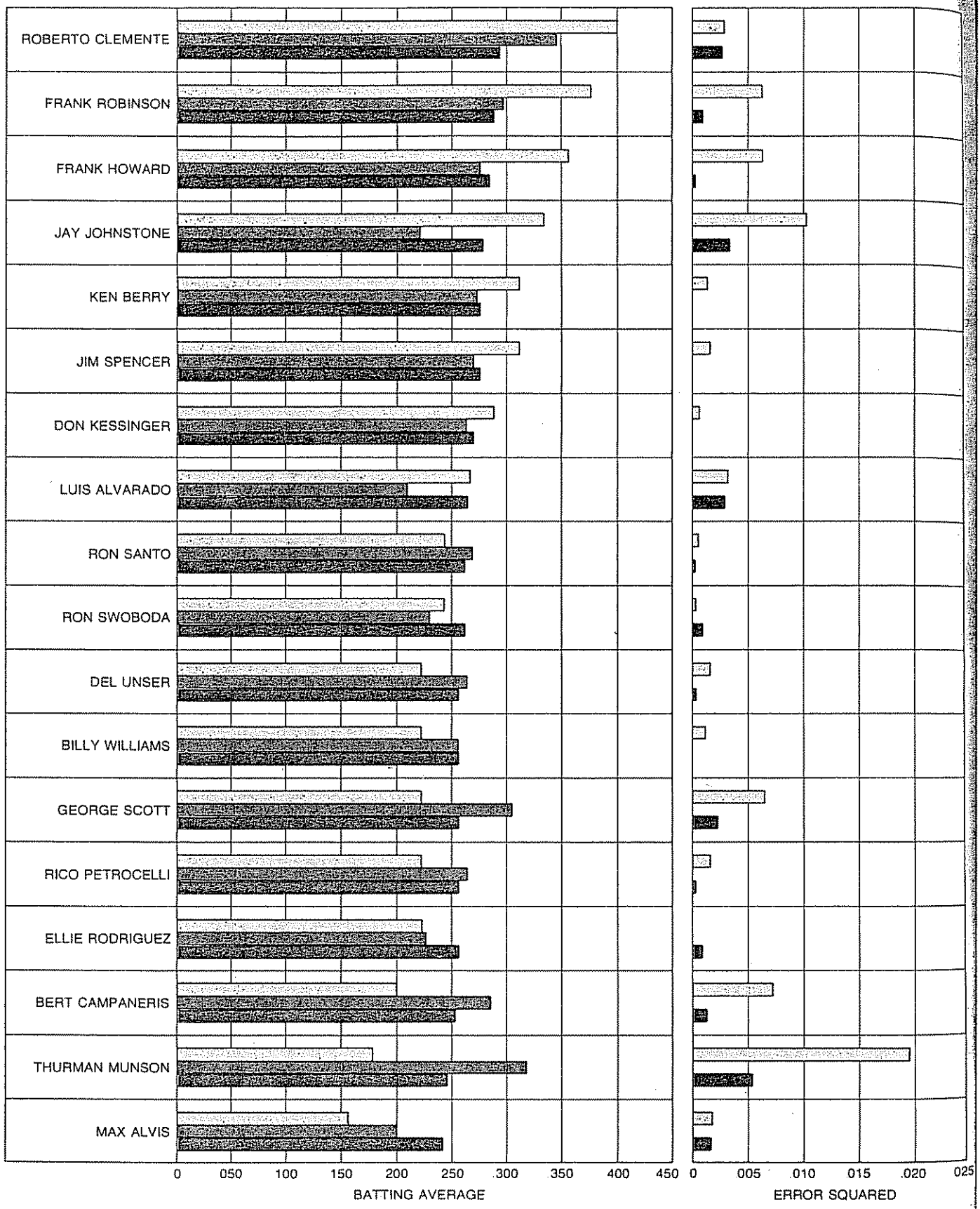
The James-Stein estimator for each player is found through the following equation: $z = \bar{y} + c(y - \bar{y})$. The quantity $(y - \bar{y})$ is the amount by which the player's batting average differs from the grand average. The equation thus states that the James-Stein estimator z differs from the grand average by this same quantity $(y - \bar{y})$ multiplied by a constant, c. The constant c is the "shrinking factor." If it were equal to 1, then the equation would state that the James-Stein estimator for a given player is identical with that player's batting average; in other words, y equals z. Stein's theorem states that the shrinking factor is always less than 1. Its actual value is determined by the collection of all the observed averages.

In the case of the baseball data, the grand average $\bar{y}$ is .265 and the shrinking factor c is .212. Substituting these values in the equation, we find that for each player z equals $.265 + .212(\bar{y} - .265)$. Because c is about .2, each average will shrink about 80 percent of the distance to the grand average, and the total spread of the averages will be reduced about 80 percent.

As an example consider the late Roberto Clemente, who was the leading batter in the major leagues when our statistics were compiled. For Clemente y is equal to .400, and z can be determined by evaluating the expression $z = .265 + .212(.400 - .265)$. The result is .294. In other words, Stein's theorem states that Clemente's true batting ability is best estimated not by .400 but lies closer to .294. Thurman Munson, in a batting slump early in the 1970 season, had an average of only .178. Substituting this value in the equation, we find that his estimated batting ability is substantially increased: the James-Stein estimator for Munson is .247.

Which set of values, y or z, is the better indicator of batting ability for the 18 players in our example? In order to answer that question in a precise way one would have to know the "true batting ability" of each player. This true average we shall designate with θ (the Greek letter theta). Actually it is an unknowable quantity, an abstraction representing the probability that a player will get a hit on any given time at bat. Although θ is unobservable, we have a good approximation to it: the subsequent performance of the batters. It is sufficient to consider just the remainder of the 1970 season, which includes about nine times as much data as the preliminary averages were based on. The expected statistical error in such a sample is small enough for us to neglect it and proceed as if the seasonal average were the "true batting ability" θ of a player. That is one reason for choosing batting averages for this example. In most problems the true value of θ cannot be determined.

One method of evaluating the two es-



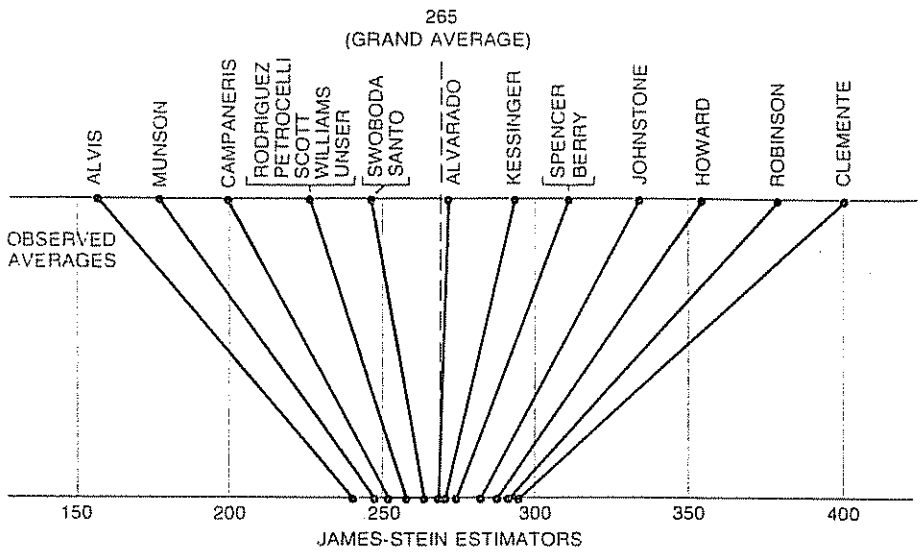
BATTING ABILITIES of 18 major-league baseball players are estimated more accurately by the method of Charles Stein and W. James than they are by the individual batting averages. The averages employed as estimators are those calculated after each player had had 45 times at bat in the 1970 season. The true batting ability of a player is an unobservable quantity, but it is closely approximated by his long-term average performance. Here the true ability is represented by the batting average maintained during the remainder of the 1970 season. For 16 of the players the initial average is inferior to another number, the James-Stein estimator, as a predictor of batting ability. The James-Stein estimators, considered as a group, also have the smaller total squared error.

timates is by simply counting their successes and failures. For 16 of the 18 players the James-Stein estimator z is closer than the observed average y to the "true," or seasonal, average θ. A more quantitative way of comparing the two techniques is through the total squared error of estimation. This is measured by first determining the actual error of each prediction, given by $(\theta - y)$ and $(\theta - z)$, for each player. Each of these quantities is then squared and the squared values are added up. The observed averages y have a total squared error of .077, whereas the squared error of the James-Stein estimators is only .022. By this comparison, then, Stein's method is 3.5 times as accurate. It can be shown that for the data given 3.5 is close to the expected ratio of the total squared errors of the two methods. We have not just been lucky.

Suppose a statistician makes a random sampling of automobiles in Chicago and finds that of the first 45 recorded nine are foreign-made and the remaining 36 are domestic. We want to estimate the true proportion of imported cars in Chicago, a quantity represented by another unobservable θ. The observed average of $9/45 = .200$ is one estimate. Another can be obtained by simply lumping this problem together with that of the 18 baseball players. Substituting the value .200 in the equation used in that problem gives a James-Stein estimator of .251 for the imported-car ratio. (Actually the addition of a 19th value changes the grand average $\bar{y}$ and also slightly alters the shrinking factor c. The changes are small, however; the amended value of z is .249.)

In this case intuition argues strongly that the observed average and not the James-Stein estimator must be the better predictor. Indeed, the entire procedure seems silly; what could batting averages have to do with imported cars? It is here that the paradoxical nature of Stein's theorem is most uncomfortably apparent. The theorem applies as well to the 19 problems as it did to the original 18. There is nothing in the statement of the theorem that requires the component problems to have some sensible relation to one another.

The same disconcerting indifference to common sense can be demonstrated in another way. What does Clemente's .400 observed average have to do with Max Alvis, who was poorest in batting among the 18 players? If Alvis had had an early-season hitting streak, batting say .444 instead of his actual .156, the James-Stein estimator for Clemente's average would have been increased from .294 to .325. Why should Alvis' success or lack of it have any influence on our estimate of Clemente's ability? (They were not even in the same league.)



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

It is questions of this kind that have been raised by critics of Stein's method. In order to reply to them it will be necessary to describe the method rather more carefully.

Taking an average is an easy and familiar process that seems to need no justification. Actually it is not obvious why the average is so often useful in estimating the true center of gravity of a random process. The explanation lies in the distribution that the values of the random variable tend to assume.

The distribution most common in scientific work is the "normal" distribution, described by a bell-shaped curve; it was first investigated in depth by Gauss and is sometimes called the Gaussian distribution. It is constructed by assuming that the random variable can take on any value along some axis; the probability that it falls within any given interval is then made equal to the area under the same interval of the bell-shaped curve. The curve is completely specified by two parameters: the mean, θ, which lies at the peak of the curve, and the standard deviation, which measures how closely the values are distributed around the mean. It is customary to assign the standard deviation the symbol σ (sigma). The larger the standard deviation is, the more widely dispersed the data are.

In probability theory a known mean and standard deviation are employed to predict future behavior. A problem in statistics proceeds in the opposite direction: from observed data the statistician must infer the mean θ and the standard deviation σ.

Suppose, for example, the measurement of some random variable x yields the five successive values 10.0, 9.4, 10.3, 8.6 and 9.7. Suppose further the values are known to be part of a normal distribution with a standard deviation of 1. What is the value of the true mean θ? In principle the mean could have any value, but some values are more likely than others. A mean of 6.5, for example, would require that all five values be under the extreme tail of the curve and that none be found near the center. Gauss showed that among all possible choices for the mean, the average $\bar{x}$ of the observed data (which in this case has a value of 9.6) maximizes the probability of obtaining the data actually seen. In this sense the average is the most likely estimate of the mean; in fact, Gauss constructed the normal distribution just so that it would have this property.

There is a further justification, also pointed out by Gauss, for choosing the average as the best estimator of the unobservable mean θ. Gauss noted that the average of the data is an "unbiased" estimator of the mean, in the sense that it favors no selected value of θ. To be more precise, the average is unbiased because the expected value of $\bar{x}$ equals the true θ no matter what θ may be. There are infinitely many unbiased estimators of θ, none of which estimates θ perfectly. Gauss showed that the expected squared error of estimation for the average $\bar{x}$ is lower than that for any other linear, unbiased function of the observations. In the 1940's it was demonstrated that no other unbiased function of the data, whether it is linear or nonlinear, can estimate θ more accurately than the average, in terms of expected squared error. An essential contribution to that proof had been made in the 1920's by

R. A. Fisher, who showed that all the information about θ that can possibly be found in the data is contained in the average $\bar{x}$.

In the 1930's a mathematically more rigorous approach to statistical inference was undertaken by Jerzy Neyman, Egon S. Pearson and Abraham Wald; the ideas they developed are part of what is now known as statistical decision theory. They discarded the requirement of unbiased estimation and examined all functions of the data that could serve as estimators of the unknown mean θ. These estimators were compared through a risk function, defined as the expected value of the squared error for every possible value of θ.

Consider three competing estimators: the average of the data, $\bar{x}$; half that average, $\bar{x}/2$, and the median of the data, or middle value. For both the average and the median the risk function is constant; that is merely another way of saying that their expected squared error in predicting the mean θ is the same no matter what the value of θ really is. Of the two constant risk functions, the one for the average $\bar{x}$ is uniformly smaller by a factor of about two-thirds; clearly the average is the preferred estimator. In the language of decision theory the median is said to be "inadmissible" as an estimator of θ, since there is another estimator that has a smaller risk (expected squared error) no matter what θ is. (It should be mentioned, however, that when the data have a distribution other than the normal one, it is possible for the order of preference to be reversed.)

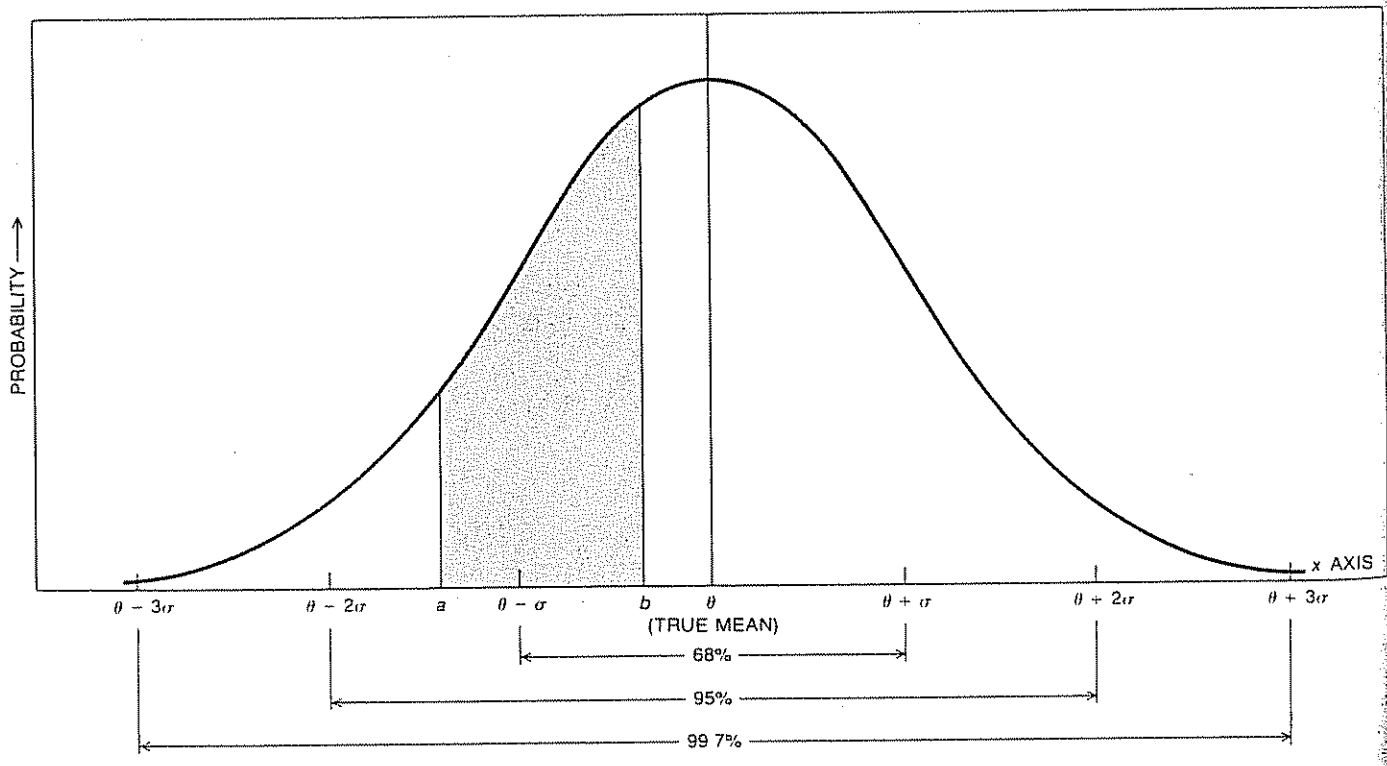
For the estimator $\bar{x}/2$, which is biased toward the value $\theta = 0$, the risk function is not constant; this estimator is accurate if θ happens be close to zero, but the expected squared error increases rapidly as the true mean departs from zero. The risk function describes a parabola, with the minimum point at $\theta = 0$; if the mean does happen to be zero, then the risk function for $\bar{x}/2$ is four times smaller than that for the average itself. At large values of the mean, however, the average $\bar{x}$ regains its superiority. With other estimators we can poke down the risk function below that of the average at any point we wish to, but it always pops up again somewhere else.

There remains the possibility that some other estimator has a risk that is uniformly lower than that of the average. In 1950 Colin R. Blyth, Erich L. Lehmann and Joseph L. Hodges, Jr., proved that no such estimator exists. In other words, the average $\bar{x}$ is admissible, at least when it is applied to one set of observations for the purpose of estimating one unknown mean.

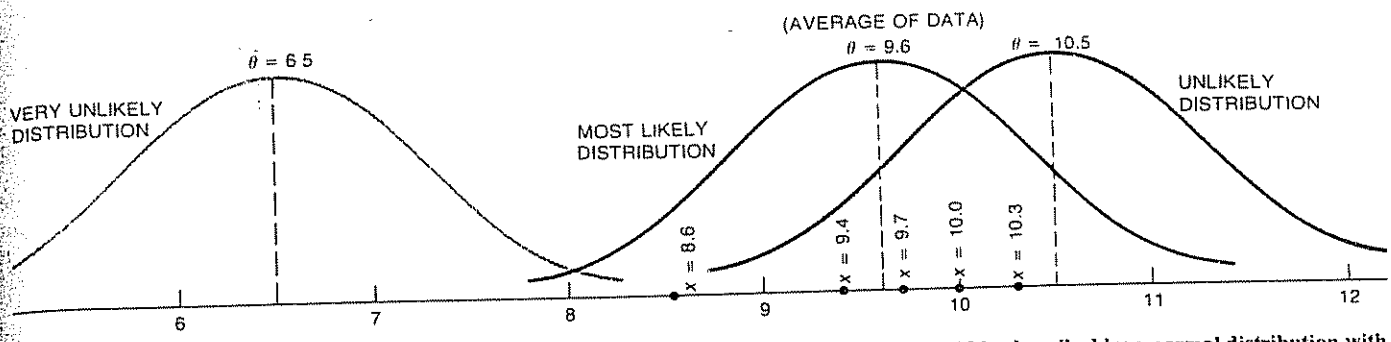
Stein's theorem is concerned with the estimation of several unknown means. No relation between the means need be assumed; they can be batting abilities or proportions of imported cars. On the other hand, the means are assumed to be independent of one another. In evaluating estimators for these means it is once again convenient to employ a risk function defined as the sum of the expected values of the squared errors of estimation for all the individual means.

The obvious first choice of an estimator for each of several means is the average of the data related to that mean. The entire historical development of statistical theory from Gauss through decision theory argues that the average is an admissible estimator as long as there is just one mean, θ, to be estimated. Stein showed in 1955 that the average is also admissible for estimating two means. Stein's paradox is simply his proof that when the number of means exceeds two, estimating each of them by its own average is an inadmissible procedure. No matter what the values of the true means, there are estimation rules with smaller total risk.

In 1955 Stein was able to prove this proposition only in those cases where the number of means, a quantity we shall designate k, was very large. Stein's 1961 paper written in collaboration with James extended the result to all values of k greater than 2; moreover, it did so in a constructive manner. Stein and James not only showed that estimators must exist that are everywhere superior to the



NORMAL DISTRIBUTION of a random variable around the mean value of that variable provides the fundamental justification for estimation by averaging. The distribution is defined by two parameters, the mean, θ, which locates the central peak of the distribution, and the standard deviation, σ, which measures how widely scattered the data points are. It is assumed in defining the distribution that the variable x can take on any value on the x axis. The most likely value of x is, by definition, the mean θ. The probability that x lies within any given interval on the axis, such as that between the points a and b, is equal to the area under the bell-shaped curve between those points.



PROBLEM IN STATISTICS is to deduce from a set of data the true mean and standard deviation of the distribution. Even when it is known that the distribution is a normal one and that the standard deviation is 1, the mean could in principle have any value. Some values, however, are more likely than others. For example, the five data points (x) given here could be described by a normal distribution with a mean of 6.5 only if all five points were more than two standard deviations above the mean. It can be shown that the data are most likely to be generated by a distribution with a mean equal to the observed average of the data, denoted $\bar{x}$. In this case the average is equal to 9.6.

averages; they were also able to provide an example of such an estimator.

The James-Stein estimator has already been defined in our investigation of batting averages. It is given by the equation $z = \bar{y} + c(y - \bar{y})$, where y is the average of a single set of data, $\bar{y}$ is the grand average of averages and c is a "shrinking factor." There are several other expressions for the James-Stein estimator, but they differ mainly in detail. All of them have in common the shrinking factor c; it is the definitive characteristic of the James-Stein estimator.

In the baseball problem c was treated as if it were a constant. Actually it is determined by the observed averages and therefore is not a constant. The shrinking factor is given by the equation

$$c = 1 - \frac{(k-3)\sigma^2}{\Sigma(y-\bar{y})^2}.$$

Here k is again the number of unknown means, σ^2 is the square of the standard deviation and $\Sigma(y-\bar{y})^2$ is the sum of the squared deviations of the individual averages y from the grand average $\bar{y}$.

Let us briefly explore the meaning of this rather forbidding equation. With k and σ^2 fixed, we find that the shrinking factor c becomes smaller (and the predicted means are more severely affected by it) as the expression $\Sigma(y-\bar{y})^2$ gets smaller. On the other hand, c increases, approaching unity, and the shrinking is less drastic as the expression $\Sigma(y-\bar{y})^2$ increases.

What do these equations mean in terms of the behavior of the estimator? In effect the James-Stein procedure makes a preliminary guess that all the unobservable means are near the grand average $\bar{y}$. If the data support that guess in the sense that the observed averages are themselves not too far from $\bar{y}$, then the estimates are all shrunk further toward the grand average. If the guess is contradicted, then not much shrinking is done. These adjustments to the shrinking factor are accomplished through the effect the distribution of averages around the grand average $\bar{y}$ has on the equation that determines c. The number of means being estimated also influences the shrinking factor, through the term $(k-3)$ appearing in this same equation. If there are many means, the equation allows the shrinking to be more drastic, since it is then less likely that variations observed represent mere random fluctuations.

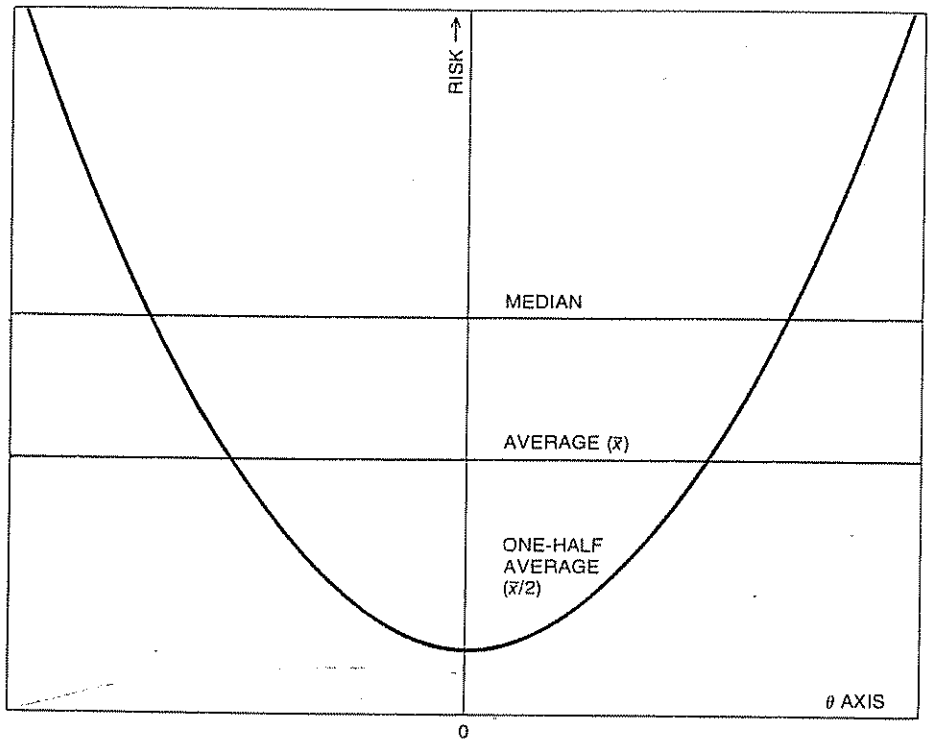
With c calculated in this manner, the risk function for the James-Stein estimator is less than that for the sample averages no matter what the true values of the means θ happen to be. The reduction of risk can be substantial, particularly when the number of means is larger than five or six. The risk function is not constant for all values of the true mean θ, as it is for the observed averages. The risk of the James-Stein estimator is smallest when all the true means are the same. As the true means depart from one another the risk of the estimator increases, approaching that of the observed averages but never quite equaling it. The James-Stein estimator does substantially better than the averages only if the true means lie near each other, so that the initial guess involved in the technique is confirmed. What is surprising is that the estimator does at least marginally better no matter what the true means are.

The expression for the James-Stein estimator that we have employed refers all observed averages to the grand average $\bar{y}$. This procedure is not the only one possible; other expressions for the estimator dispense with $\bar{y}$ entirely. What cannot be avoided is the introduction of some more or less arbitrary initial guess or point of origin for the estimator. The observed averages, it will be noted, do not depend on a choice of origin. Before Stein discovered his method it was felt that such "invariant" estimators must be preferable to those whose predictions change with each choice of an origin. The theory of invariance, to which Stein had been a principal contributor, was badly shaken by the James-Stein counterexample. From the standpoint of mathematics this is the most unsettling aspect of Stein's theorem. Indeed, the paradox was not discovered earlier largely because of a strong prejudice that the estimation problem, being stated without reference to any particular origin, should be solved in a similar way.

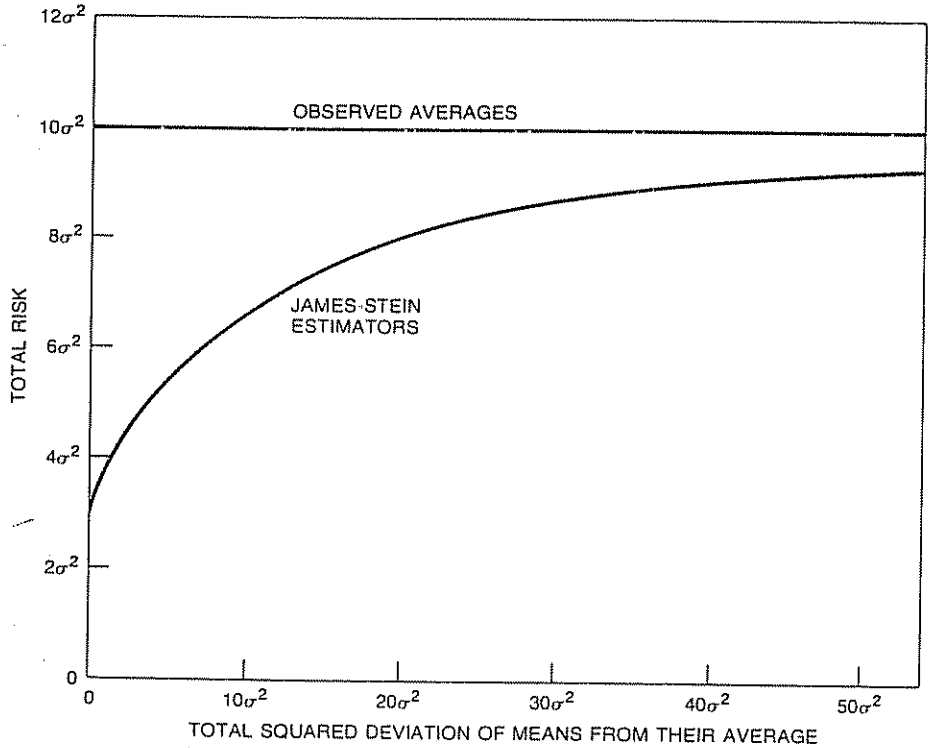
Applications of Stein's method tend to involve large sets of data with many unknown parameters. Some of the difficulties of such problems, as well as the practical potential of the method itself, can be illustrated by an example: an analysis of the distribution of the disease toxoplasmosis in the Central American country of El Salvador.

Toxoplasmosis is a disease of the blood that is endemic in much of Central America and in other regions of the Tropics. In El Salvador roughly 5,000 people drawn in varying numbers from 36 cities were tested for toxoplasmosis. The observed rate of incidence for each city can conveniently be expressed by comparison with the national rate (that is, with the grand average $\bar{y}$). A measured rate of .050, for example, denotes a city with an incidence of the disease 5 percent higher than the national average. The measured rates have an approximately normal distribution. The standard deviations of these distributions are known, but they differ from city to city, depending inversely on how large a sample population was tested in that city. It is the task of the statistician to estimate the true mean θ of the distribution for each city from the measured incidence y.

In this case the appropriate form of the James-Stein estimator is $z = cy$. The simplification, which was introduced by us, is made possible by the chosen manner of expressing the observations y. They are defined in such a way that the grand average $\bar{y}$ is zero, and terms containing $\bar{y}$ therefore drop out of the equation. On the other hand, the estimation



VARIOUS ESTIMATORS of a single true mean, θ, can be evaluated by way of a risk function. The risk is defined as the expected value of the squared error of estimation, considered as a function of the mean θ. The average of the data, $\bar{x}$, is an estimator with a constant risk function: no matter what the true mean is, the expected value of the squared error is the same. The median, or middle value, of the data also has constant risk, but it is everywhere greater (by a factor of 1.57) than the risk of the average. Half the average ($\bar{x}/2$) is an estimator whose risk depends on the actual value of the mean; the risk is smallest when the mean is near zero and increases rapidly when the mean departs from zero. For the estimation of a single mean there is no estimator with a risk function that is everywhere less than the risk function of the average $\bar{x}$.



TOTAL RISK FUNCTION for the James-Stein estimators is everywhere less than that for the individual observed averages, as long as the number of means being estimated is greater than two. In this case there are 10 unknown means. The risk is smallest when all the means are clustered at a single point. As the means depart from one another the risk of the James-Stein estimators increases, approaching that of the observed averages but never quite reaching it.

procedure is now complicated by the fact that the shrinking factor c is different for each city, varying inversely as the standard deviation of y for that city. This dependence of the shrinking factor on the standard deviation has a simple intuitive rationale. A large standard deviation implies a high degree of randomness or uncertainty in a measurement. If the measured incidence is unusually large, it can therefore be attributed more reasonably to random fluctuations within the normal distribution than to a genuinely large value of the true mean θ. It is thus proper to reduce this value drastically, that is, to apply a small shrinking factor.

The same argument can be made even more forcefully by returning for a moment to baseball. Frank O'Connor pitched for Philadelphia in the 1893 season. He batted twice in his major-league career, hitting successfully both times. His observed batting average is hence 1.000. The James-Stein rule for the 18 players considered above estimates O'Connor's true batting ability to be $.265 + .212(1.000 - .265) = .421$ (ignoring the effect of the new data on the grand average and on the shrinking factor). This is a silly estimate, although not as silly as 1.000. A perfect average after two times at bat is not at all inconsistent with a true value in the range from .242 to .294 that is estimated for the other players. The shrinking constant c applied to O'Connor's average should be severer in order to compensate for the smaller amount of data available for him.

For the El Salvador observations, most of the shrinking factors are quite gentle, between .6 and .9, but a few are in the range from .1 to .3. Which set of numbers should we prefer, the James-Stein estimators or the measured rates of incidence? That depends largely on what we want to use the numbers for.

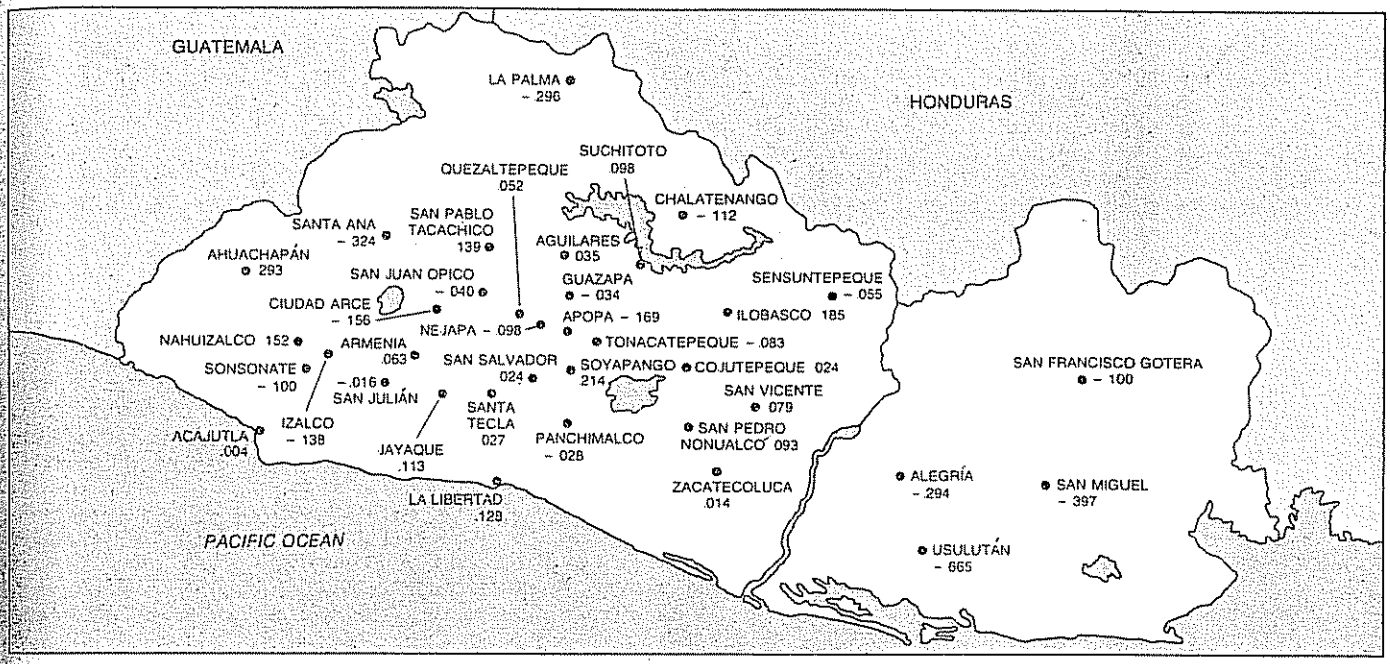
If the Minister of Health for El Salvador intends to build local hospitals for people suffering from toxoplasmosis, the James-Stein estimators probably offer the more reliable guidance. The reason is that the expected value of the total squared error is smaller for the James-Stein estimators; in fact, it is smaller by a factor of about three. The important point in this calculation is that the expected error is added up for all the cities. Any particular hospital might be the wrong size or in the wrong place, but the sum of all such mismatches would be smaller for the James-Stein estimators than for the observed rates.

The James-Stein estimators are also likely to be preferable for determining the ordering of the true means. In this regard it is notable that the city with the highest apparent incidence (according to the measured rates y) is ranked 12th according to the James-Stein estimators

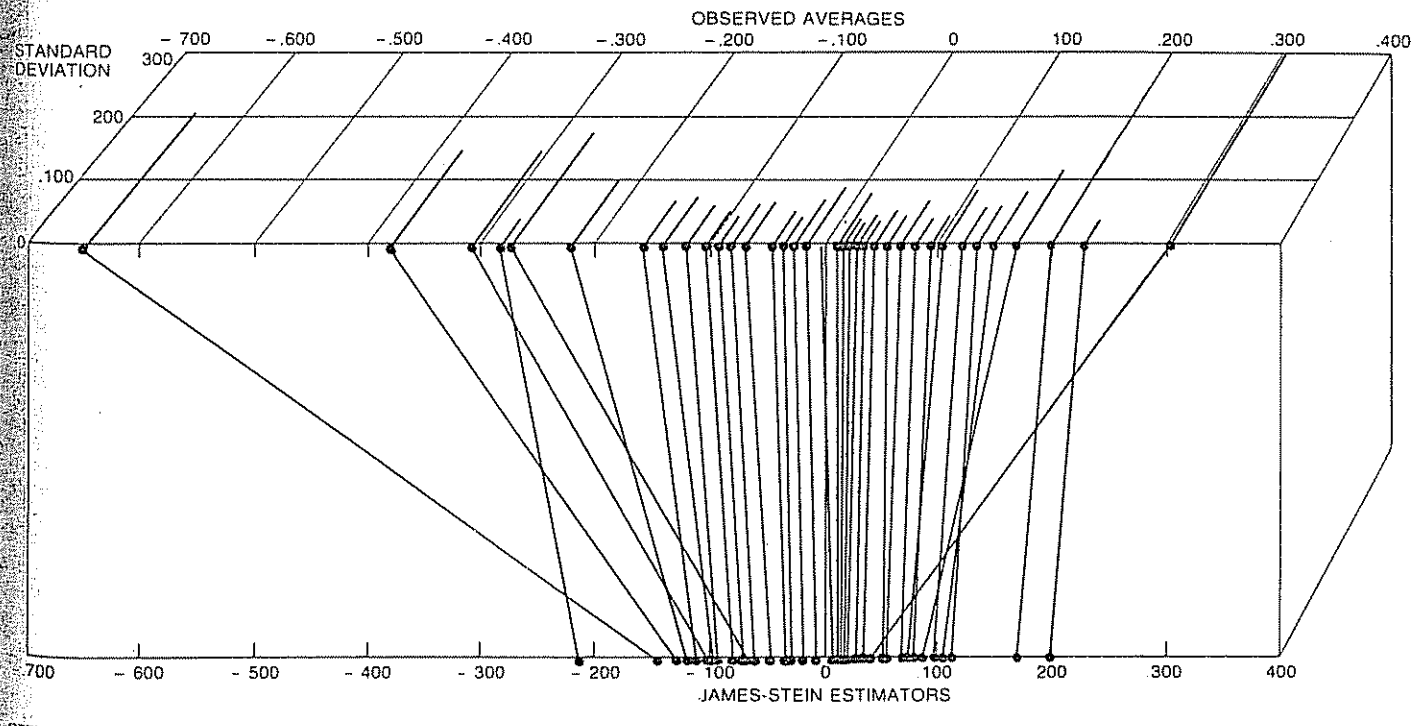
The estimate is drastically reduced because the sample was very small in that city. This information might be useful if there were funds for only one hospital.

Suppose an epidemiologist wants to investigate the correlation of the true incidence in each city with attributes such as rainfall, temperature, elevation or population? Once again the James-Stein estimators are preferred; a rough calculation shows that they would give a closer approximation in about 70 percent of the cases.

There is one purpose for which the measured incidence may well be superior to the James-Stein estimator: when a single city is considered in isolation. As we have seen, the James-Stein method gives better estimates for a majority of cities, and it reduces the total error of estimation for the sum of all cities. It



INCIDENCE OF TOXOPLASMOSIS, a disease of the blood, was surveyed in 36 cities in the Central American country El Salvador. The measured incidence in each city can be regarded as an estimator of the true incidence, which is unobservable. The measured incidence has a normal distribution whose standard deviation is determined by the number of people surveyed in that city. The measured rates are expressed in terms of deviation from the national incidence (the average of the rates observed in all the cities). Thus zero denotes exactly the national rate, and a city with a measured incidence of −.040 would have an observed rate 4 percent lower than the country as a whole.



SHRINKING of the observed toxoplasmosis rates to yield a set of James-Stein estimators substantially alters the apparent distribution of the disease. The shrinking factor is not the same for all the cities but instead depends on the standard deviation of the rate measured in that city. A large standard deviation implies that a measurement is based on a small sample and is subject to large random fluctuations; that measurement is therefore compressed more than the others are. In the El Salvador data the most extreme observations tend to be correlated with the largest standard deviations, again suggesting the unreliability of those measurements. Compared with the observed rates, the James-Stein estimators can be proved to have a smaller total error of estimation. They also provide a more accurate ranking of the cities.

cannot be demonstrated. however, that Stein's method is superior for any particular city; in fact. the James-Stein prediction can be substantially worse

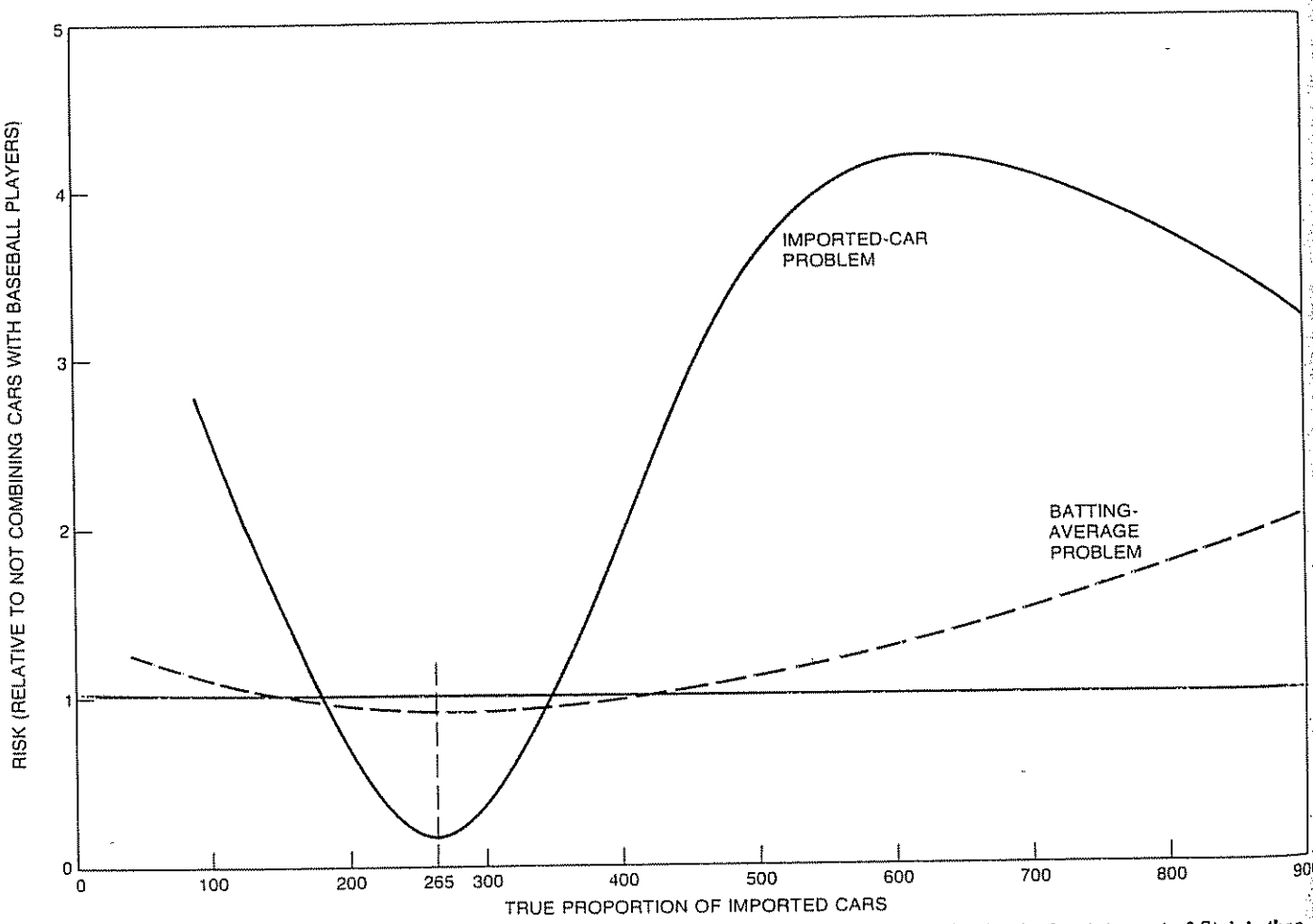
Estimating the true mean for an isolated city by Stein's method creates serious errors when that mean has an atypical value The rationale of the method is to reduce the overall risk by assuming that the true means are more similar to one another than the observed data. That assumption can degrade the estimation of a genuinely atypical mean. Now we see why imported cars should not be included in the same calculations with the 18 baseball players. There is a substantial probability that the automobiles will be atypical.

Suppose we ignore this hazard and lump together all 19 problems; we can then calculate the total expected squared error as a function of the true percentage of imported cars. It turns out that the risk for both the baseball players and the automobiles is reduced only if the percentage of imported cars happens to lie in the same range as the estimated batting averages; otherwise the risk of error for both kinds of problem is increased.

The question of whether or not a particular mean is "typical" is a subtle one whose implications are not yet fully understood. Returning to the problem of toxoplasmosis in El Salvador, let us single out for attention the city of Alegría, which has the fifth-smallest measured incidence of the disease: −.294. It is one of four cities included in the survey that are east of the Rio Lempa; all four have distinctly negative values of measured incidence y. It is plausible to suppose that this is no coincidence and that the rate of toxoplasmosis east of the Lempa is genuinely lower. A James-Stein estimator that consolidates information from the entire country therefore may be less than optimal in these cities. We have developed techniques for taking advantage of extra information of this kind. but the theory underlying those techniques remains rudimentary.

An astute follower of baseball might be aware that just as each player's batting ability can be represented by a Gaussian curve. so too the true batting abilities of all major-league players have an approximately normal distribution. This distribution has a mean of .270 and a standard deviation of .015. With this valuable extra information. which statisticians call a "prior distribution," it is possible to construct a superior estimate of each player's true batting ability. This new estimator, which we shall give the label Z, is defined by the equation $Z = m + C(y - m)$. Here y is again the observed batting average of the player, but $\bar{y}$, the grand average, has been replaced by m, the mean of the prior distribution. which is known to have the value .270. In addition there is a different shrinking factor, C, which depends in a simple way on the standard deviation of the prior distribution (equal to .015).

This procedure is not a refinement of Stein's method; on the contrary. it predates Stein's method by 200 years. It is the mathematical expression of a theorem published (posthumously) in 1763 by the Reverend Thomas Bayes.



UNRELATED PROBLEMS can be lumped together for analysis by Stein's method, but only at the risk of increasing error. To the 18 batting averages computed earlier, for example, one might add a 19th number representing the proportion of imported cars observed in Chicago. New James-Stein estimators could then be calculated for both the baseball players and the automobiles, based on the grand average of all 19 numbers. Nothing in the statement of Stein's theorem prohibits such a procedure, but the evident illogic of it has justifiably been criticized. In fact, including the unrelated data can reduce the risk function only if the proportion of imported cars happens to be near the mean batting average of .265; otherwise the expected error of estimation for both the cars and the baseball players is increased.

He was able to show that this estimator minimizes the expected squared error associated with the randomness in both the observed averages (y) and in the true means (θ).

The formula for the James-Stein estimator is strikingly similar to that of Bayes's equation. Indeed, as the number of means being averaged grows very large, the two equations become identical. The two shrinking factors c and C converge on the same value, and the grand average $\bar{y}$ becomes equal to the mean m precisely when all players are included in the calculation. The James-Stein procedure, however, has one important advantage over Bayes's method. The James-Stein estimator can be employed without knowledge of the prior distribution; indeed, one need not even suppose the means being estimated are normally distributed. On the other hand, ignorance has a price, which must be paid in reduced accuracy of estimation. We have shown that the James-Stein method increases the risk function by an amount proportional to $3/k$, where k is again the number of means being estimated. The additional risk is therefore negligible when k is greater than 15 or 20, and it is tolerable for k as small as 9.

In this historical context the James-Stein estimator can be regarded as an "empirical Bayes rule," a term coined by Herbert E. Robbins of Columbia University. In work begun in about 1951 Robbins demonstrated that it is possible to achieve the same minimum risk associated with Bayes's rule without knowledge of the prior distribution, as long as the number of means being estimated is very large. Robbins' theory was immediately recognized as a fundamental breakthrough; Stein's result, which is closely related, has been much slower in gaining acceptance.

The James-Stein estimator is not the only one that is known to be better than the sample averages. Indeed, the James-Stein estimator is itself inadmissible! Its failure lies in the fact that the shrinking factor c can assume negative values, and it then pulls the means away from the grand average rather than toward it. When that happens, simply replacing c with zero produces a better estimator. This estimator in turn is also inadmissible, but no uniformly better estimator has yet been found.

The search for new estimators continues. Recent efforts have been concentrated on achieving results like those obtained with Stein's method for problems involving distributions other than the normal distribution. Several lines of work, including Stein's and Robbins' and more formal Bayesian methods, seem to be converging on a powerful general theory of parameter estimation.