# MODERN STATISTICS AND THE COMPUTER REVOLUTION

Jerome H. Friedman, Professor
Computation Research Group, Stanford Linear Accelerator Center

## INTRODUCTION OF DR. FRIEDMAN

Dr. Pyke: In the statistical sciences, the growth of knowledge is accompanied by the growth of data. A few years ago there was a young person giving a talk at a meeting in this city on a new method of analyzing very large sets of complex data. That person is our next speaker, Dr. Jerome Friedman. Jerry had created a work station at the Stanford Linear Accelerator Center where this analysis could be done. It's a bit of a shock to realize that this was less than ten years ago, yet there were at the time only one or two other work stations in the country capable of doing this. The changes since then have been quite dramatic.

Dr. Friedman, like our previous speaker, holds a Ph.D. in the area of theoretical physics. His experience with experimentation and data from experiments led him to statistics, and he saw the needs there and the potential for great innovation. Dr. Friedman is currently at Stanford University. He is a member of the Committee on Applied and Theoretical Statistics (CATS) of the Board on Mathematical Sciences and consults with many laboratories on the analysis of large data sets. Jerry's talk today, on "Modern Statistics and the Computer Revolution," will concern not just computer-aided analysis of data, but also the role of the computer in the collection of data.

Dr. Friedman: Thank you very much, Ron. I want also to thank CATS and the Board on Mathematical Sciences for inviting me. It is a real honor to be asked to speak here today.

The first thing I would like to mention is that this is a personal view. I am not sure that I represent statistics or even academic statistics in my views of how modern statistics is responding to the computer revolution. I want you to keep in mind that these are my particular views, and I was asked to be somewhat controversial. So I will try.

The first controversial statement I am going to make is that statistics is about data. What does that mean? It is the basic science that tries to tell us how to take the data provided by our senses and transform it into valid information. We have five senses, and there are extensions -- instruments from microscopes to telescopes to particle accelerators to sample surveys.

*"The first controversial statement I am going to make is that statistics is about data."*

We want to use that information or that data to provide valid inferences about our environment, the systems under study, and to try to do that in the most powerful ways possible. Most important, and this is what separates statistics to a large degree from other information sciences, is that we seek to understand the limits of the validity of the inferences. The previous speaker talked a lot about this, and statisticians do this more than anyone else. Unfortunately, often the public and policy-makers don't like to be told about the limits of the validity of the inferences.

> *"What separates statistics to a large degree from other information sciences is that we seek to understand the limits of the validity of the inferences."*

After I left physics, I worked in computer science before I became a statistician. My former colleagues in both physics and computer science often ask what is so fascinating about statistics that I should leave two very interesting fields to work in statistics. My reason is that statistics is the basic information science. It is the one that tells us how to take the information our senses gather and make valid inferences.

If you believe that, and if you want to understand how the computer revolution is changing statistics, we need to consider how it is changing data. How is the computer changing the way that data are collected, why they are collected, and what we want the data to tell us? Those are three things we have to look at, and if we understand them, we will have some motivation for the philosophy behind the changes that are being made in statistics.

## HOW DATA ARE COLLECTED

Let us look at how data are collected. Computers have allowed us to collect data automatically. This happened in some sciences well before others; in my science, high-energy physics, this had already begun in the late sixties. Now it is very common in many settings. Not all data, of course, are taken automatically, but now a great deal is. Some examples include remotely sensed satellite data, weather monitoring, and air pollution monitoring. In the latter case, you have a station with many sensors that are sampling the air once every second, sometimes ten times a second, making lots of measurements and writing them to a computer.

> *"Computers have allowed us to collect data automatically."*

Other examples include data base transactions, like point of sale at a grocery store when your product is wiped over a bar code reader, that is recording a lot of data to be stored about the nature of the sales. Or, when someone asks a credit data base about your credit, that request is also stored and becomes part of the data record.

Computers are now frequently used to control industrial automation. We heard about that from Dr. Frosch. They control laboratory instrumentation as well, and when you have computers controlling processes, you can also have the computers automatically logging and recording data. Of course, you can also do computer simulations. It is very popular now to simulate traffic patterns, to simulate all kinds of systems on a computer. Taking data then amounts to putting the right statements in the computer program running the simulation to collect data at various points of the simulation.

Now, how are these data different from the kind of data that we have seen in the past? First of all, they are untouched by human hands. That is both good and bad. The good news is that we tend to have fewer transcription errors. The bad news, and it really is bad news, is that fewer and far less sophisticated checks are being made on it. When you handle each datum by hand (especially if the handler is a scientist himself or a trained technician) you get to know the name of every piece of data that goes by -- you can see it immediately. You can use your cognitive resources to tell you whether that piece of data makes sense. Even sometimes when you haven't formally written down what it means for the data to make sense, you just know that it cannot be right. The computer cannot know that. You can put all kinds of checks in the computer, but from the point of reference of human reasoning they are very simple indeed. So, we have many fewer checks for reasonableness.

> *"The good news is that we tend to have fewer transcription errors.... The bad news is that fewer and far less sophisticated checks are being made on it."*

Another thing that is very different with automatically collected data is the cost structure. Systems for automatically collecting data are very expensive to set up. You have to buy the computer and buy the sensing equipment. You have to write the software, and you have to connect it up and make it work. That is costly and usually takes a lot of time, but once you have done that, the marginal cost for taking data is just the price of the magnetic medium upon which you record it.

In order to amortize this very large setup cost, people tend to collect massive amounts of data in this setting, both in terms of number of replications or observations and also the number of measurements per observation. If you are going to set this thing up, you might as well set it up to collect a lot of data. That is a big difference from hand-gathered data.

## WHY DATA ARE COLLECTED

Next the question of why data are collected; what are we asking data to do for us that we didn't ask before? And how are computers changing that? I think it is fair to say that we ask data to do far more for us than ever before. I have some examples of this, and they by no means exhaust the possibilities. These are some of the things I either know about or have been involved in.

*"I think it is fair to say that we ask data to do far more for us than ever before."*

Forecasting is a big business. People use data to forecast. The reason, of course, is clear because he or she who knows the future makes a great deal of money. Consequently people use data from the past (hoping the system isn't changing too much) to forecast what is going to happen in the future. Weather prediction and medical diagnosis are well-known examples. A third example, which governments use heavily, is demand for goods and services. Another is securities prices; all Wall Street firms have large statistical packages that try to use current securities prices and other economic conditions to predict future securities prices. Of course, a well-known use is tax audits. The Internal Revenue Service gets millions of tax returns. They audit some of them. Some of those audits yield a lot of return; some don't yield very much. They can take that data to build a model to tell them what a return with a particular set of characteristics will yield. The situation with loan applications is very similar. Banks have a huge data base of people who have asked for loans, people they have given loans to, and those who have defaulted and those who have not. They want to build a model based on that data to predict who is going to default in the future.

There are other areas where people use data in a different way now. For example, pattern recognition is a big business, from printed characters to automated reading of handwriting to trying to recognize isolated and continuous speech. Surveillance is another big business. We heard from the last speaker about process and quality control. Computers control assembly lines or at least control the collection of data which are then used to control the assembly line. Often optimization of a process is the goal of this data collection.

These are some of the things data is being used for now that I don't think were necessarily envisioned by people many years ago when many of our current statistical procedures were developed.

## WHAT DO WE WANT DATA TO TELL US?

Now to the third question -- what do we want data to tell us? I think this is far less focused now than in the past because of automated data collection. When data are expensive to obtain, that is, they have to be taken by hand, you don't really collect any more than is necessary to answer the question you have in mind. However, when data are marginally cheap, you tend to collect massive amounts of data on the off chance that it might be useful someday.

*"What do we want data to tell us? I think this is far less focused now than in the past because of automated data collection.... When data are marginally cheap, you tend to collect massive amounts of data on the off chance that it might be useful someday."*

You don't want somebody to stop you and say, "Gee, if you had only taken this piece of data, we could have answered the question." So, you tend to take everything you can see, and this is one of the symptoms of the so-called "information explosion." People are taking lots and lots of data without giving a great deal of thought to how they expect it to be used. There is nothing wrong with that; it just means that we have to develop methodology that deals with that situation rather than situations where the data is collected for well-focused and specific purposes.

Now, it is a fact of life that nearly all statistical methods that are in common use today were developed before about 1950. You might say, "1960." You might say, "1940," but certainly, they were developed before computers were widely used. Of course, they were invented by very clever people, and they were intended to take maximal advantage of the situation at that time.

*"It is a fact of life that nearly all statistical methods that are in common use today were developed before about 1950."*

## STATISTICS IN THE PRE-COMPUTER ERA

What was the situation then: small data sets. Why? Because it was expensive to collect data. You had a small number of measurements that you made for each observation, and you tended to take a small number of observations. Also, the earliest users of statistics were in agriculture, psychology, sociology, and medicine. The data tended to come from noisy environments, unlike engineering and the physical sciences where the noise is a less dominant part of the measurement process.

In those days, in fact, even when I was a young physicist, the thought was that if you had to use statistical methods to analyze an experiment, you had not done the right experiment -- the answer ought to be obvious if you made the measurements well enough. Physicists no longer believe that. Chemists, I think, no longer believe that either, and so, these scientists, and engineers as well, are tending to use statistical methods more. Of course, at first they had to do manual computation. In those days, if you wanted to do a linear least squares fit, which is now considered a trivial computation, you had a room full of Monroe calculators and you went to work with people cranking the Monroes to try to do a linear least squares fit. Similar efforts were needed to get statistical tables for confidence intervals or whatever else you wanted to do.

So, this was the setting -- small data sets, manual computation, and noisy environments. These were the conditions under which almost all the statistical procedures that we use today were produced.

What were the consequences of these constraints? Basically, the statistical procedures had to be simple. They had to be computationally simple if they were going to be used and the analysis of their properties and performance had to be mathematically tractable. The only way you could figure out how well a statistical procedure would do -- how small the uncertainty would be after you applied the procedure -- was to do it mathematically. Clearly, that was a major limitation. As a result, advances in statistics paralleled advances in mathematics or mathematics applied to statistics.

> *"The only way you could figure out how well a statistical procedure would do -- how small the uncertainty would be after you applied the procedure -- was to do it mathematically. Clearly, that was a major limitation."*

But the simplicity required restrictions. You had to answer well-focused questions, like tests of hypotheses. Is there an effect or isn't there? You had to make strong assumptions about the setting from which the data came; about the relationships among the variables, for instance, that the relationship is linear or additive; and about the error mechanisms, e.g., that errors are normal or Poisson, usually independent and identically distributed. These were strong assumptions and generally unverifiable, so you rarely knew if they were justified or not. (You may have wondered sometimes how two sets of researchers could take the same data and come to different conclusions. They can do it by making different assumptions about the mechanisms that underlie the system.) Some assumptions were necessary in order to keep the procedures mathematically tractable, more precisely in order to permit analysis of the performance properties of the procedures, and it had to be done with mathematics.

Now, I think it is fair to say that despite these limitations, statistics has evolved into a spectacularly successful information science, and this is a great tribute to statisticians in the twenties, thirties, and forties, and even going back to Gauss and Laplace. In a time well before the computer and information age, they invented procedures that are still highly useful today. So, you might say, "That is a tough act to follow. Is there anything that one can do to make things better?"

## ENTER THE COMPUTER

First of all, let us look at the response of statistics when computers did become available. The first response of the field of statistics to computers was the so-called "statistical package." The idea was to take these useful methods that were, of course, invented in the pre-computer age, program them, collect them on the same file, and include rudimentary data management facilities and a simple language for invoking them.

*"The first response of the field of statistics to computers was the so-called 'statistical package.'"*

The packages made it far more convenient to utilize the precomputer techniques, and this approach was highly successful. Companies like SAS, BMDP, and MINITAB make a lot of money doing exactly this, and to this day this is the principal response of statistics to computing. Whenever a new personal computer comes out, an IBM, Macintosh, or whatever, one of the first sets of software after the text editor and the spreadsheet is the statistics package, and these statistical packages are essentially the same as the statistical packages that were developed in the sixties as the initial response of statistics to computing.

It turns out that truly computer-oriented statistical methods are just now beginning to emerge, about 20 years after computing became generally available. That is rather remarkable, given, as I tried to indicate before, the dramatic change in the nature of the data we are seeing. The obvious question seems to be, "How can computers fundamentally affect statistical methodology, as opposed to simply implementing the old methodology on a computer instead of a Monroe calculator?"

*"It turns out that truly computer-oriented statistical methods are just now beginning to emerge, about 20 years after computing became generally available."*

## COMPUTER-DRIVEN STATISTICAL METHODOLOGY

There are three underlying factors. First of all, with computers, it is possible to develop statistical methodologies such that tractability of calculations is no longer a requirement. Second, mathematical tractability need no longer be a requirement. Third, exploratory data analysis is now possible. That is largely a consequence of larger data sets; less well-focused questions can now be asked. The paradigm of classical statistics required not looking at the data before formulating a question. Then you asked the data, "Yes or no? Does it support or not support the hypothesis?" Often analysts didn't look at the data simply to see what it revealed, in other words to look for the unexpected; unexpected observations like outliers, unexpected relationships between the variables, things we just didn't anticipate.

*"The paradigm of classical statistics required not looking at the data before formulating a question. Then you asked the data, 'Yes or no? Does it support or not support the hypothesis?'"*

Computers allow us to do exploratory data analysis, to make many pictures of the data, and to do many analyses of the data in an attempt to discover the unexpected. So, these are the three key changes: computational and mathematical tractability are no longer requirements, and the computer permits us to explore data. I will take up each of these in turn.

*"Computers allow us to do exploratory data analysis, to make many pictures of the data, and to do many analyses of the data in an attempt to discover the unexpected."*

Computational Tractability

Before computers, our estimates, or statistical procedures which resulted in estimators, and our summary statistics had to be represented in closed-form mathematical formulas. A formula is a simple prescription for computation, one that does not contain data dependent branches. The optimum solution in statistics is generally expressed as a maximum likelihood or least squares estimate, that is, the estimate that does "best" in this particular situation where "best" is determined by a likelihood function or a least squares criterion. These solutions had to be explicitly obtained in terms of simple mathematical prescriptions, namely formulas, and that was quite a limitation. If you had a procedure that seemed intuitively reasonable but was too hard to compute, or if your estimator was a likelihood solution that couldn't be evaluated in closed form, you were out of luck.

Now, however, computers have changed all that. Because of the tremendous computational power that computers provide, we can now inexpensively calculate very complex estimators. We don't require solutions in closed form. We can use iterative numerical methods and numerical optimization to great advantage.

*"Because of the tremendous computational power that computers provide, we can now inexpensively calculate very complex estimators. We don't require solutions in closed form."*

We can say that the maximum likelihood estimate is the number or the set of numbers that maximizes a function and let a computer find that maximum, without ever having to write in closed form what that maximum is. We can describe our procedures not in the language of formulas but in the language of algorithms and data structures, the basic languages of computer programming. This allows us to fit far more flexible models with many fewer assumptions concerning their structure. Many researchers are now trying to fit more generalized nonlinear models in regression settings, for example, using intensive computing with many less assumptions.

## Mathematical Tractability

It may not be obvious how computers can help with mathematical tractability. Before we had computers, the basic tool was mathematics founded in probability theory, and it was the key to understanding our statistical procedures, in particular how they varied with sampling. This included computing standard errors and so-called "significance," calculating power, and demonstrating optimality, i.e., that a certain procedure gives the most precise estimate. Thus mathematics was the principal tool, and advances in statistics were predicated on clever advances in mathematics. This required procedures that used the data in a very simple way. The procedure did not look at the data and decide what to do with it; it gave a simple prescription for doing something straightaway. You had to assume simple probability models, such as Gaussian or Poisson. You had to assume idealized settings; for instance, that the sample size was infinite when, of course, it is never infinite, although it may be getting bigger all the time. You had to make unverifiable assumptions. You had to make a set of assumptions, find your estimates, and hope those assumptions characterized the system. Or you could do a sensitivity study and try different sets of assumptions to see if it made a lot of difference; often, it did.

With computers, I think the most dramatic advance has been in this area of mathematical tractability, in the development of simulation and sample reuse.

> *"With computers, I think the most dramatic advance has been in this area of mathematical tractability, in the development of simulation and sample reuse."*

## Simulation and Sample Reuse

These two techniques relieve the mathematical burden and allow us to be more imaginative with our procedures, without worrying so much about how mathematically tractable they are.

With computer simulation, you can not only simulate traffic patterns on a freeway, but you can simulate sampling in statistical procedures. So, if you have any crazy procedure that you invent that goes to work on the data, and you assign any probability model you like, you can use computer simulation to obtain the sampling distribution, standard errors, confidence intervals, etc., and compare it with other methods. In fact, you can perform complete simulation experiments to try to identify the best methods in any particular situation. One of the earliest and most successful examples of this was in the early seventies, the so-called "Princeton robustness study,"[1] where the investigators considered, I think, well over 100, perhaps 200 estimators of location, and they simulated the estimators' performance in a variety of settings so that they could make recommendations on what estimators to use in what context. That was a pure simulation study, with very little mathematics.

In this sense statistics, at least academic statistics, is becoming an experimental science, as well as a branch of mathematics. You can either try to analyze your procedure mathematically or set up a simulation experiment to try to analyze it. Both are powerful ways to do it.

> *"In this sense statistics....is becoming an experimental science, as well as a branch of mathematics."*

Finally, the most exciting development, I think, is in the sample reuse techniques, because they relieve us of having to make assumptions, unverifiable assumptions. With the sample reuse techniques, you can again use any crazy procedure that anyone invents to apply to the data at hand. You don't specify a probability model. In some sense, you would like the data to tell you the probability model from which it was generated. You can use cross validation to assess future performance of models, that is, use part of the data itself in the absence of future data to assess future performance.

> *"The most exciting development, I think, is in sample reuse techniques, because they really relieve us of having to make assumptions, unverifiable assumptions."*

It is a whole talk on how you do that, and you can use so-called "jackknifing and bootstrapping" techniques to estimate the sampling distributions to obtain confidence intervals and standard errors. What these techniques do is resample from the data as if the data were the population. That is why one of these methods is called the bootstrap, because you seem to be pulling yourself up by your bootstraps. With these techniques one can show that those sampling distributions you get are very

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). Robust Estimators of Location: Survey and Advances. Princeton, N.J.: Princeton U. Press.

often quite close to sampling distributions from the real population, even though you don't know what the distribution of the real population is. You only have to assume independence; you don't need to have any specific probability model for the error mechanism.

## EXPLORATORY DATA ANALYSIS

I will say a little bit about data exploration and then summarize. I think of data exploration as searching for the unexpected. This is something that I think the inventors of statistical techniques in the thirties and forties didn't think much about, because then the emphasis was in another direction. Analysts looked at their data and disagreed (sometimes strenuously) on the results, whether the data said "Yes" or "No." Hypothesis testing was a real intellectual triumph; everyone could agree on a procedure to be applied to the data to determine the probability of yes and the probability of no.

### Interactive Computer Graphics

Now, the main development in computing that has allowed us to do data exploration has been advances in interactive computer graphics. Any numerical procedure, no matter how inventive, must specify the effects it attempts to model. On the other hand, by definition something isn't unexpected if you can write a program to detect it.

> *"The main development in computing that has allowed us to do data exploration has been advances in interactive computer graphics."*

The human gift for pattern recognition is very powerful at seeing undefined effects. You can look and be surprised at something that you never dreamed would occur. So, the idea here is to make a series of informative pictures of the data in the most powerful way to uncover the unanticipated. You look at the data and seek to figure out what they are trying to tell you without formulating any well-focused questions at all. That is the stuff that Nobel prizes are made of, and of course, it is always wise to explore the data before modeling it because whenever you explore the data, you always find surprises. You may have in mind a particular set of assumptions to apply to some data. When you explore it, you may see that those assumptions make no sense whatsoever.

> *"The idea here is to make a series of informative pictures of the data in the most powerful way to uncover the unanticipated."*

Now, most data that one collects are multidimensional or multivariate. The goal of this kind of exploratory data analysis is to map this high-dimensional data (data where you have measured many things) to lower, humanly perceivable dimensionality, where you can look at the data and apply the human gift for pattern recognition.

This research has two parts. One part is to try to map the data from the high dimension to the lower dimension in a way that reduces the information content as little as possible, and the other is to raise humanly perceivable dimensionality as high as possible, to try to make them meet. Whenever you reduce dimensionality, you lose some information, but some mappings to lower dimensional spaces lose less information than others. Here I am going to talk about the second part, that is, trying to increase the humanly perceivable dimensionality, and that you do with interactive color graphics. With graphics alone you can make two-dimensional views of your data, scatter plots or scatter diagrams. With color graphics you can add a third diffuse viewing dimension. On an ordinary scatter plot you can locate a point probably to one part in a hundred. When you try to use color hues to tell you the value of a third variable, you can see perhaps three or four distinctly.

The big change here is individual computing power. The new work stations allow mammoth amounts of computing power to be dedicated to an individual user.

Now, this has allowed us to put another dimension into our plots, namely the dimension of time. If we can recompute pictures (meaning different maps of the high-dimensional data to lower dimensional settings) ten to 30 times a second, we can get the appearance of continuous motion, and the simplest example of that is if you consider a three-dimensional point cloud viewed on a two-dimensional screen. If you subject it to a rigid rotation, your brain transforms the three dimensions of horizontal, vertical, and time to horizontal, vertical, and depth. It just happens automatically in the circuitry of your brain, and you see it in three dimensions almost immediately. That is the simplest application of using time as another dimension to show plots.

The other method is interaction. These maps from the high-dimensional setting to the low-dimensional setting are always controlled by certain parameters. In an interactive graphics setting, the value of these parameters can be changed by graphics input devices, like joysticks, tracker balls, or mice, and so you can have something analogous to video games where you are changing the mapping. The picture is changing for you in real time, and that gives you derivative information. You move the tracker ball or the mouse, the picture changes, and you may see that things got worse. You back off, move in another direction, and see that things get a little better. The process continues as long as the analyst wishes. We have several movies of the research that we have done in this area, trying to illustrate someone at a console doing this kind of interactive data exploration.

## SUMMARY

What can one conclude about the computer's role in modern statistics? I think it is fair to say that computers are providing both a need for new statistical methodology and, of course, the ability to accomplish it. The need arises from much larger data sets taken from less noisy environments, and far more demanding applications than we have ever had before. Of course, it provides us the ability as well because we now have these computer-intensive methods that allow us to perform thousands to millions of computations on each individual observation -- something unthinkable before computers -- not ten computations or 100 computations, but sometimes millions of computations on each number, each datum that we have collected.

> *"Computers are providing both a need for new statistical methodology and the ability to accomplish it.... The need arises from much larger data sets taken from less noisy environments, and far more demanding applications than we have ever had before."*

I think what this allows us to do, and what is basically the trend for the future, is that we are substituting computer power for unverifiable assumptions about the data. That is the goal of these computer-intensive methods.

> *"What this allows us to do, and what is basically the trend for the future, is that we are substituting computer power for unverifiable assumptions about the data."*

In estimation and modeling they allow us to use complex estimators like the robust estimators in nonlinear fitting schemes. In inference, there are the sample reuse techniques which are massively computationally intensive: cross validation, jackknife, and the bootstrap. They allow us now to explore data because you can do things like the very computationally intensive projection pursuit, as well as construct very rapidly a large number of pictures to view.

Why use these techniques? I think the reason is clear. The cost of computation is ever decreasing, but the price that we pay for the incorrect assumptions is staying the same.

> *"The cost of computation is ever decreasing, but the price we pay for the incorrect assumptions is staying the same."*

## DISCUSSION

Professor Cohen (ROCKEFELLER UNIVERSITY): How widely diffused among actual working statisticians are the computer-intensive techniques that you described today?

Dr. Friedman: Not very, and that is because they are all fairly new. I recall an occasion when Sam Karlin (Stanford University) was describing to me a procedure that he was applying, and he said, "You probably haven't heard of it. It's brand new -- it's only 15 years old." These procedures aren't 15 years old yet, and so they don't see widespread use. Their use will grow, but there has not yet been a great body of experience with them. I think by far the most rapidly advancing is the bootstrap as a measure of variability, and that is moving very fast in terms of acceptance.

> *"I think by far the most rapidly advancing [method in terms of use] is the bootstrap as a measure of variability, and that is moving very fast in terms of acceptance."*

Dr. Tufte (YALE UNIVERSITY): Could you give us some kind of assessment of computer graphics and computer statistical systems, perhaps by name for, say, mainframes and down to work station size?

Dr. Friedman: I have not actually made a market survey in the last year of commercially available graphics systems, either software or hardware. So anything I would say would be dated, and I would rather not get into comparing vendors and which brand names I like and don't like. Different people have different tastes. I will say that I really do like the Macintosh and especially the new Mac II, the color Mac.

Dr. Pyke: There is, of course, tremendous concern today about the quality and standards of those software packages.

Dr. Eddy (CARNEGIE-MELLON UNIVERSITY): I wanted to make a comment about the past successes of statistical packages and the future failures of statistical packages. Previously statistics has been successful in the broad scientific research community because packages produce numbers like .05 and smaller, such numbers which could then be used as evidence of the success of whatever particular theory was being tested. The future techniques that Dr. Friedman has talked about involve subjective, interactive assessment of the data and do not allow for objective assessment by such statistical procedures.

> *"Previously statistics has been successful in the broad scientific research community because packages produce numbers like .05.... The future techniques that Dr. Friedman has talked about involve subjective, interactive assessment of the data and do not allow for objective assessment by such statistical procedures."*

Consequently, I think you are going to discover that the major commercial vendors of software are going to be extremely reluctant to adopt these techniques. I am sure that there will be new upstart companies to distribute them to those of us who enjoy playing with these things, but I am very pessimistic about the future adoption of these very useful and powerful techniques in commercially available software. I would be interested to hear your comments.

Dr. Friedman: You may be right. I am not sure that I agree with your thesis that the success of the packages is due to the fact that they "give an objective answer." First of all, they don't give an objective answer. Every procedure that you invoke has a set of assumptions associated with it that you are tacitly agreeing to which may or may not be valid, and you can apply several procedures in the package to the same set of data and get different results.

Many of the new computational intensive procedures allow, in principle, calculation of confidence intervals and significance tests, as I tried to indicate with the bootstrap. There I showed you a picture, but I didn't have to. One could at each point have located the 95 percent interval, or whatever you want. I presume what you are talking about is the exploratory techniques.

> *"A lot of the new computational intensive procedures allow, in principle, calculation of confidence intervals and significance tests."*

Dr. Eddy: Let me give you a specific example. The Princeton Robustness Study was published in 1972, and there were some specific recommendations from that study that we can infer by reading it carefully. There is not one single major statistical package that has implemented those estimation procedures as a routine part of its analysis. There is at least one package which prints out some of the numbers in a page with 50,000 other numbers, but this is not a routinely accepted part of statistical analysis. That is my concern, 15 years later.

Dr. Friedman: I understand that, and it is a surprise to me that robust methods have not been adopted by traditional statisticians, and they have been around for at least 15 years, unlike some of the things I have talked about. A reason for that was given to me by Werner Stuetzle (University of Washington). He says that they don't do something fundamentally different or new. They are estimating location, and you had techniques for estimating location before, and now you have a slightly better technique for estimating location. Is it really worth implementing this for the slight improvement you may or may not get, as opposed to some other computer-intensive methods which give you fundamentally new capabilities with real promise of success? I was probably a little pessimistic in my response to Dr. Cohen. Some of these methods are seeing a lot of use, especially in industry where people are betting money on the outcome. I agree with you that in academia where the goal is to get a paper published, to prove that the effect is a real one, they won't be accepted too readily.

> *"Some of these methods are seeing a lot of use, especially in industry where people are betting money on the outcome."*

But in industry where people bet money on the outcome, where it is important to get the right answer, I think if you can demonstrate that you are getting better answers, people will use them.

Dr. Eddy: The purpose of my raising this entire issue was to urge you and your fellow researchers in this area to develop procedures which allow the publication of numbers like .05.

Dr. Friedman: We are trying.