Bayesians, Frequentists, and Scientists

Bradley EFRON

Broadly speaking, nineteenth century statistics was Bayesian, while the twentieth century was frequentist, at least from the point of view of most scientific practitioners. Here in the twenty-first century scientists are bringing statisticians much bigger problems to solve, often comprising millions of data points and thousands of parameters. Which statistical philosophy will dominate practice? My guess, backed up with some recent examples, is that a combination of Bayesian and frequentist ideas will be needed to deal with our increasingly intense scientific environment. This will be a challenging period for statisticians, both applied and theoretical, but it also opens the opportunity for a new golden age, rivaling that of Fisher, Neyman, and the other giants of the early 1900s. What follows is the text of the 164th ASA presidential address, delivered at the awards ceremony in Toronto on August 10, 2004.

KEY WORDS: Bootstrap; Empirical Bayes; False discovery rate; Simultaneous testing.

Two Septembers ago, there was a conference of particle physicists and statisticians at Stanford, called phystat2003. I gave a talk at phystat2003 titled "Bayesians, Frequentists, and Physicists." Earlier that year I'd spoken to a meeting of biomedical researchers at the "Hutch" in Seattle, under the title "Bayesians, Frequentists, and Microbiologists." These weren't the same lectures, and both were different than tonight's talk, but you can see that I've gotten stuck on a naming scheme. You might worry that this has gotten out of hand, and next week it might be almost anything else that ends in "ists." But no, there is a plan here: The common theme I've been stuck on, and what I want to talk about tonight, is the impact of statistics on modern science and also the impact of modern science on statistics.

Statisticians, by the nature of our profession, tend to be critical thinkers, and that includes a big dose of self-criticism. It is easy to think of statistics as a small struggling field, but that's not at all what the historical record shows. Starting from just about zero in 1900, statistics has grown steadily in numbers and, more importantly, in intellectual influence. The growth process has accelerated in the past few decades as science has moved into areas where random noise is endemic and efficient inference is crucial.

It's hard to imagine phystat1903, back when physicists scorned statistical methods as appropriate only for soft noisy fields like the social sciences. But physicists have their own problems with noise these days, as they try to answer questions where data are really thin on the ground. The example of greatest interest at phystat2003 concerned the mass of the neutrino, a famously elusive particle that is much lighter than an electron and may weigh almost nothing at all.

The physicists' trouble was that the best unbiased estimate of the neutrino mass was negative, about -1 on a scale with unit standard error. The mass itself can't be negative of course, and these days they're pretty sure it's not zero. They wished to establish an upper bound for the mass, the smaller the better from the point of view of further experimentation. As a result, the particle physics literature now contains a healthy debate on Bayesian versus frequentist ways of setting the bound. The current favorite is a likelihood ratio-based system of one-sided confidence intervals.

The physicists I talked with were really bothered by our 250-year-old Bayesian–frequentist argument. Basically, there's only one way of doing physics, but there seems to be at least two ways to do statistics, and they don't always give the same answers. This says something about the special nature of our field. Most scientists study some aspect of nature: rocks, stars, particles. We study scientists, or at least scientific data. Statistics is an information science, the first and most fully developed information science. Maybe it's not surprising then that there is more than one way to think about an abstract subject like "information."

The Bayesian–frequentist debate reflects two different attitudes about the process of doing science, both quite legitimate. Bayesian statistics is well suited to individual researchers, or a research group, trying to use all of the information at its disposal to make the quickest possible progress. In pursuing progress, Bayesians tend to be aggressive and optimistic with their modeling assumptions. Frequentist statisticians are more cautious and defensive. One definition says that a frequentist is a Bayesian trying to do well, or at least not too badly, against any possible prior distribution. The frequentist aims for universally acceptable conclusions, ones that will stand up to adversarial scrutiny. The FDA, for example, doesn't care about Pfizer's prior opinion of how well it's new drug will work, it wants objective proof. Pfizer, on the other hand, may care very much about its own opinions in planning future drug development.

Bayesians excel at combining information from different sources, "coherence" being the technical word for correct combination. On the other hand, a common frequentist tactic is to pull problems apart, focusing, for the sake of objectivity, on a subset of the data that can be analyzed optimally. I'll give examples of both tactics soon.

Broadly speaking, Bayesian statistics dominated nineteenth century statistical practice, while the twentieth century was more frequentist. What's going to happen in the twenty-first century? One thing that's already happening is that scientists are bringing statisticians much bigger datasets to analyze, with millions of data points and thousands of parameters to consider all at once. Microarrays, the thing I was talking about to the microbiologists in Seattle, are the poster boy for scientific giganticism.

Classical statistics was fashioned for small problems, a few hundred data points at most and a few parameters. Some new thinking is definitely called for on our part. I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian

B. Efron is Professor of Statistics and Biostatistics at Stanford University. This research was supported by NIH grant 8R01 EB002784 and NSF grant DMS-00-72360.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Presidential Address
DOI 10.1198/016214505000000033

and frequentist reasoning. Tonight I'm going to argue that in some ways, huge datasets are actually easier to handle for both schools of thought.

Here's a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union, she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked whether the doctor had given her any more information. "Yes," she said, "he told me that the proportion of identical twins was one-third." This is the population proportion, of course, and my friend wanted to know the probability that *her* twins would be identical.

Bayes would have lived in vain if I didn't answer my friend using Bayes's rule. According to the doctor, the prior odds ratio of identical twins to nonidentical twins is one-third to two-thirds, or one-half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of identical twins. Bayes's rule says to multiply the prior odds by the likelihood ratio to get the current odds; in this case $1/2$ times 2 equals 1, or in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50–50 (wishing the answer had come out something else, like 63–37, to make me seem more clever). Incidentally, the twins are a couple of years old now and "couldn't be more nonidentical," according to their mom.

Now Bayes rule is a very attractive way of reasoning, and fun to use, but using Bayes rule doesn't make one a Bayesian. *Always* using Bayes rule does, and that's where the practical difficulties begin. The kind of expert opinion that gave us the prior odds one-third to two-thirds usually doesn't exist, or may be controversial or even wrong. The likelihood ratio can cause troubles too. Typically the numerator is easy enough, being the probability of seeing the data at hand given our theory of interest, but the denominator refers to probabilities under other theories, which may not be clearly defined in our minds. This is why Bayesians have to be such aggressive math modelers. Frequentism took center stage in the twentieth century to avoid all of this model specification.

Figure 1 concerns a more typical scientific inference problem, of the sort that is almost always handled frequentistically these days. It involves a breast cancer study that attracted national attention when it appeared in the *New England Journal of Medicine* in 2001. Dr. Hedenfalk and his associates were studying two genetic mutations that each lead to increased breast cancer risk, called BRCA1 and BRCA2 by geneticists. These are different mutations on different chromesomes. Hedenfalk et al. (2001) wondered whether the tumors resulting from the two different mutations were themselves genetically different.

- BRCA1 (7 Tumors)

-1.29 -1.41 -0.55 -1.04 1.28 -0.27 -0.57

- BRACA2 (8 Tumors)

-0.70 1.33 1.14 4.67 0.21 0.65 1.02 0.16

Figure 1. Expression Data for the First of 3,226 Genes, Microarray Study of Breast Cancer (Hedenfalk et al. 2001).

To answer this question, Hedenfalk et al. (2001) took tumor material from 15 breast cancer patients, 7 from women with the BRCA1 mutation and 8 from women with BRCA2. A separate microarray was developed for each of the 15 tumors, with each microarray having the same 3,226 genes. Here we see the data only for the first gene: seven genetic activity numbers for the BRCA1 cases and eight activity numbers for the BRCA2 cases. These numbers don't have much meaning individually, even for microbiologists, but they can be compared with each other statistically. The question of interest is whether the expression levels are different for BRCA1 and BRCA2. It looks like this gene might be more active in the BRCA2 tumors, because those eight numbers are mostly positive, whereas six of the seven BRAC1s are negative.

A standard frequentist answer to this question uses Wilcoxon's nonparametric two-sample test (which amounts to the usual t-test except with ranks replacing the original numbers). We order the 15 expression values from smallest to largest and compute "W," the sum of ranks for the BRCA2 values. The biggest W could be is 92, if all eight BRCA2 numbers were larger than all seven BRCA1s; at the opposite end of the scale, if the eight BRCA2s were all smaller than the seven BRCA1s, we'd get $W = 36$. For the gene 1 data, we actually get $W = 83$, which looks pretty big. It *is* big by the usual frequentist criterion. Its two-sided p value, the probability of getting a W at least this extreme, is only .024 under the null hypothesis that there is no real expression difference. We'd usually put a star next to .024 to indicate significance, according to Fisher's famous .05 cutoff point. Notice that this analysis requires very little from the statistician; no prior probabilities or likelihoods, and only the specification of a null hypothesis. It's no wonder that hypothesis testing is wildly popular with scientists, and has been for 100 years.

The .05 significance cutoff has been used literally millions of times since Fisher proposed it in the early 1900s. It has become a standard of objective comparison in all areas of science. I don't think that .05 could stand up to such intense use if it wasn't producing basically correct scientific inferences most of the time. But .05 was intended to apply to a single comparison, not 3,226 comparisons at once.

I computed W for each of the 3,226 genes in the BRCA microarray data. The histogram in Figure 2 shows the results, which range from eight genes with the smallest possible W, $W = 36$, to seven genes with $W = 92$, the largest possible, and with all intermediate values represented many times over. (There's more about the analysis of this data set in Efron 2004.)

It looks like something is definitely going on here. The histogram is much wider than the theoretical Wilcoxon null density (the smooth curve) that would apply if none of the genes behaved differently for BRCA1 and BRCA2. A total of 580 of these genes (18% of them) achieve significance according to the usual one-at-a-time .05 criterion. That's a lot more than the null hypothesis 5%, but now it isn't so clear how to assess significance for any one gene given so many candidates. Does the $W = 83$ that we saw for gene 1 really indicate significance?

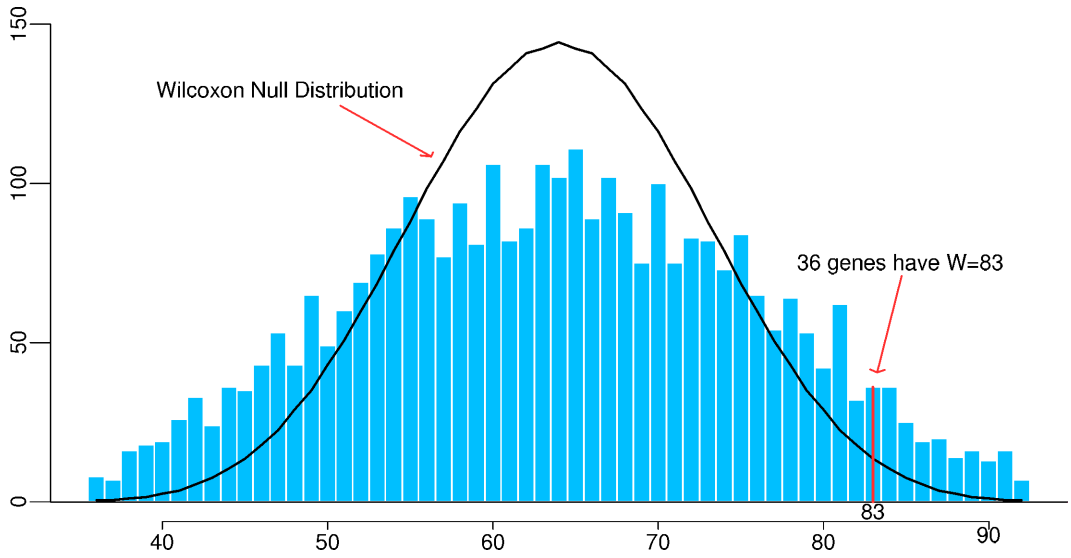

Figure 2. Wilcoxon Statistics for 3,226 Genes From a Breast Cancer Study.

I was saying earlier that huge datasets are in some ways easier to analyze than the small ones we've been used to. Here is a big-dataset kind of answer to assessing the significance of the observed Wilcoxon value $W = 83$. A total of 36 of the 3,226 genes (gene 1 and 35 others) have $W = 83$; under the null hypothesis that there's no real difference between BRCA1 and BRCA2 expression, we would expect to see only 9 genes with $W = 83$. Therefore, the expected false discovery rate is 9 out of 36, or 25%. If Hedenfalk decides to investigate gene 1 further, he has a 25% chance of wasting his time. Investigator time is usually precious, so he might prefer to focus attention on those genes with more extreme W values having smaller false discovery rates. For example, there are 13 genes with $W = 90$, and these have a false discovery rate of only 8%.

The "9 out of 36" calculation looks definitely frequentist. In fact, the original false discovery rate theory developed by Benjamini and Hochberg in 1995 was phrased entirely in frequentist terms, not very much different philosophically than Fisher's .05 cutoff or Neyman–Pearson testing. Their work is a good example of the kind of "new theory" that I hope statisticians will be developing in response to the challenge of massive datasets.

It turns out that the false discovery rate calculations also have a very nice Bayesian rationale. We assume that a priori a proportion p_0 of the genes are null, and that these genes have W's following the null Wilcoxon density $f_0(w)$. In a usual microarray experiment, we'd expect most of the genes to be null, with p_0 no smaller than, say, 90%. The remainder of the genes are nonnull and follow some other density, let's call it $f_1(w)$, for their Wilcoxon scores. These are the "interesting genes," the ones that we want to identify and report back to the investigators. If we know p_0, f_0, and f_1, then Bayes's rule tells us right away what the probability is of a gene being null or nonnull given its Wilcoxon score W.

The catch is that to actually carry out Bayes's rule, we need to know the prior quantities p_0, $f_0(w)$, and $f_1(w)$. This looks pretty hopeless without an alarming amount of prior modeling and guesswork. But an interesting thing happens with a large dataset like this one: We can use the data to estimate the prior quantities, then use these estimates to approximate Bayes rule. When we do so, the answer turns out much the same as before, for example, null probability 9 out of 36 given $W = 83$.

This is properly called an "empirical Bayes" approach. Empirical Bayes estimates combine the two statistical philosophies; the prior quantities are estimated frequentistically to carry out Bayesian calculations. Empirical Bayes analysis goes back to Robbins and Stein in the 1950s, but they were way ahead of their time. The kind of massively parallel datasets that really benefit from empirical Bayes analysis seem to be much more a twenty-first century phenomenon.

The BRCA dataset is big by classical standards, but it is big in an interesting way; it repeats the same "small" data structure again and again, so we are presented with 3,226 similar two-sample comparisons. This kind of parallel structure gives the statistician a terrific advantage, just what we need to bring empirical Bayes methods to bear. Statisticians are not passive observers of the scientific scene. The fact that we can successfully analyze ANOVA problems leads scientists to plan their experiments in ANOVA style. In the same way we can influence the design of big datasets by demonstrating impressively successful analyses of parallel structures.

We have a natural advantage here. It's a lot easier to manufacture high-throughput devices if they have a parallel design. The familiar medical breakthrough story on TV, showing what looks like a hundred eyedroppers squirting at once, illustrates parallel design in action. Microarrays, flow cytometry, proteomics, time-of-flight spectroscopy all refer to machines of this sort that are going to provide us with huge datasets nicely suited for empirical Bayes methods.

Figure 3 shows another example. It concerns an experiment comparing seven normal children with seven dyslexic kids. A diffusion tensor imaging scan (related to fMRI scanning) was done for each child, providing measurements of activity at 16,000 locations in the brain. At each of these locations, a two-sample t-test was performed comparing the normal and dyslexic kids. The figure shows the signs of the t-statistics

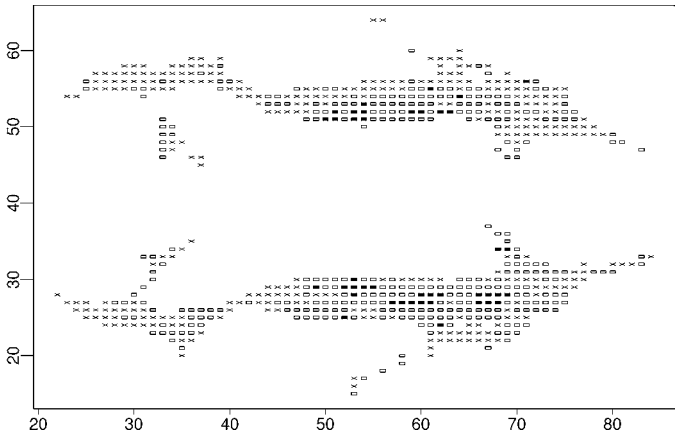

Figure 3. A Total of 580 t-Statistics From a Brain Imaging Study of Dyslexia (solid squares $t \geq 2$; empty squares $t \geq 0$; x's $t < 0$). From data of Schwartzman, Dougherty, and Taylor (2004).

for 580 of the positions on a single horizontal slice of the brain scan. (There are 40 other slices with similar pictures.) Squares indicate positive t-statistics and x's indicate negative t-statistics, with filled-in squares indicating values exceeding 2; these are positions that would be considered significantly different between the two groups by the standard .05 one-at-a-time criterion.

We can use the same false discovery rate empirical Bayes analysis here, with one important difference: The geometry of the brain scan lets us see the large amount of spatial correlation. Better results are obtained by averaging the data over small contiguous blocks of brain position—better in the sense of giving more cases with small false discovery rates. The best way of doing so is one of those interesting questions raised by the new technology.

There's one last thing to say about my false discovery rate calculations for the BRCA data: They may not be right! At first glance, the "9 out of 36 equals 25% false discoveries" argument looks too simple to be wrong. The 9 in the numerator, which comes from Wilcoxon's null hypothesis distribution, is the only place where any theory is involved. But that's where potential trouble lies. If we only had data for one gene, say for gene 1 as before, then we would *have* to use the Wilcoxon null, but with thousands of genes to consider at once, most of which are probably null, we can empirically estimate the null distribution itself. Doing so gives far fewer significant genes in this case (as you can read about in Efron 2004). Estimating the null hypothesis itself from the data sounds a little crazy, but that's what I meant about huge datasets presenting new opportunities as well as difficulties.

I have to apologize for going on so long about empirical Bayes, which has always been one of my favorite topics, and now at last seems to be going from ugly duckling to swan in the world of statistical applications. Here is another example of Bayesian–frequentist convergence, equally dear to my heart.

Figure 4 tells the unhappy story of how people's kidneys get worse as they grow older. The 157 dots represent 157 healthy volunteers, with the horizontal axis their age and the vertical axis a measure of total kidney function. I've used the "lowess" curve fitter, a complicated sort of robust moving average, to

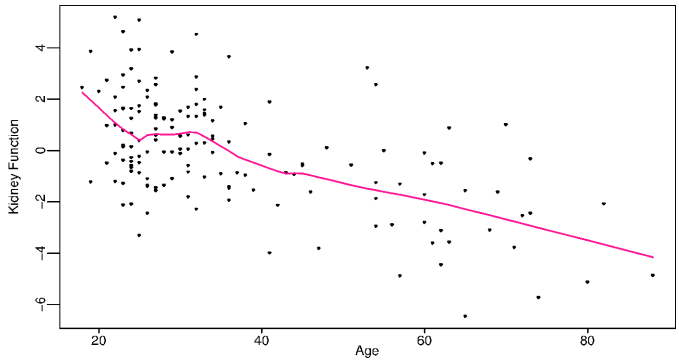

Figure 4. Kidney Function versus Age for 157 Normal Volunteers, and Lowess Fit.

summarize the decline of kidney function with age. The fitted curve goes steadily downward except for a plateau in the 20s.

How accurate is the lowess fit? This is one of those questions whose answer has gone from hopeless to easy with the advent of high-speed computation. A simple bootstrap analysis gives the answer in literally seconds. We resample the 157 points, that is, take a random sample of 157 points with replacement from the original 157 (so some of the original points appear once, twice, three times, or more and others don't appear at all in the resample). Then the lowess curve fitter is applied to the resampled dataset, giving a bootstrap version of the original curve.

In Figure 5 I've repeated the whole process 100 times, yielding 100 bootstrap lowess curves. Their spread gives a quick and dependable picture of the statistical variability in the original curve. For instance, we can see that the variability is much greater near the high end of the age scale, at the far right, than it is in the plateau.

The bootstrap was originally developed as a purely frequentist device. Nevertheless, the bootstrap picture has a Bayesian interpretation: If we could put an "uninformative" prior on the collection of possible age-kidney curves, that is, a prior that reflects a lack of specific opinions, then the resulting Bayes analysis would tend to agree with the bootstrap distribution. The bootstrap-objective Bayes relationship was pursued by Efron and Tibshirani (1998).

This brings up an important trend in Bayesian statistics. Objectivity is one of the principal reasons that frequentism dominated twentieth-century applications; a frequentist method like Wilcoxon's test, which is completely devoid of prior opinion,

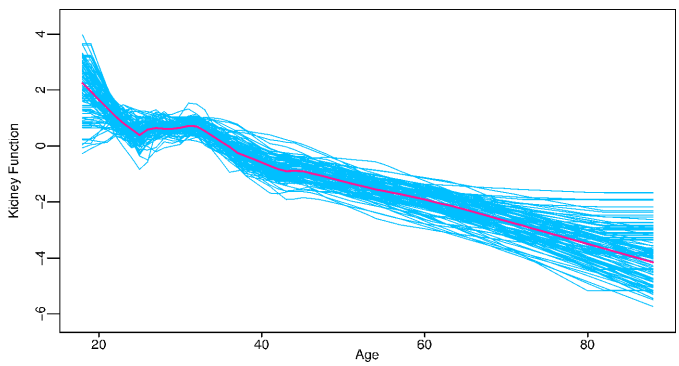

Figure 5. 100 Bootstrap Replication of Lowess Fit.

has a clear claim to being objective—a crucial fact when scientists communicate with their skeptical colleagues. Uninformative priors, the kind that also have a claim to objectivity, are the Bayesian response. Bayesian statistics has seen a strong movement away from subjectivity and toward objective uninformative priors in the past 20 years.

Technical improvements, the computer implementation of Markov chain Monte Carlo methods, have facilitated this trend, but the main impetus, I believe, is the desire to compete with frequentism in the domain of real-world applications. Whatever the reason, the effect has been to bring Bayesian and frequentist practice closer together.

In practice, it isn't easy to specify an uninformative prior, especially in messy-looking problems like choosing a possibly jagged regression curve. What looks uninformative enough often turns out to subtly force answers in one direction or another. The bootstrap connection is intriguing, because it suggest a simple way of carrying out a genuinely objective Bayesian analysis, but this is only a suggestion so far.

Perhaps I've let my enthusiasm for empirical Bayes and the bootstrap run away with the main point I started out to make. The bottom line is that we have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of classical statistics. In the struggle to find these answers, the statistics profession needs to use both frequenstist and Bayesian ideas, as well as new combinations of the two. Moreover, I think this is already beginning to happen... which was the real point of my examples.

A field like statistics has both an inside and an outside. The outside part faces our clients, the people who need answers to pressing statistical questions. My examples tonight concerned outside relationships with physicists, microbiologists, and brain researchers. One of the encouraging trends in statistics has been our increasing engagement with front-line science. This came first to industry and government groups, and now is sweeping the universities. It has changed statistics department faculties, the type of students entering the field, the problems we work on, and the articles in our journals. The change is definitely for the better—we are a much healthier profession now than when I was a student in the 1960s.

I find the microarray story particularly encouraging. First of all, biologists did come to us for answers to their difficult new inference problems. This is our reward for being helpful colleagues in the past, with all those ANOVA, *t*-tests, and randomized clinical trials that have become a standard part of biomedical research. Second, statisticians have made a serious effort to again be of help, with some of us (although not me, I'm afraid) devoting enormous energy to learning the biological-medical background of microarray technology. Most important, we actually *have* been of help. There has been definite progress made on microarray inference (a very small part of which I discussed this evening), with lots more on the way, I hope.

Microbiologists talk with other information scientists too, such as data miners, neural networkers, and bioinformatics people. It's human nature to worry about competition like this. In fact, however, we have a positive regression coefficient with these "rival" fields. Their enthusiastic energy is refreshing and contagious. They bring new data-analytic ideas into our field, ideas that statisticians can then understand and explain in terms of basic inferential theory. Many scientists are excellent probabilists, but in my experience only statisticians are trained in the kind of reverse thinking, from observed data back to possible models, necessary for inference. In other words, don't worry about statistics going out of business from outside competition.

If you do feel the need to worry, a better subject is our own production of useful new ideas. This relates to the "inside" of the statistics profession, the side that worries about the structure of statistical inference and how it can be extended. New ideas are the coin of the realm for an intellectual discipline. Without them a field hollows out, no matter how successful it may be in terms of funding or public recognition. Too much "inside" can be deadly for a field, cutting it off from the bigger world of science, as happened to mathematics in the twentieth century. Statistics had an inside phase itself in the 1950s and 1960s, but that is definitely not today's problem. In fact, I would give statistics at least passing grades for the production of genuinely useful new ideas, like empirical Bayes and false discovery rates, and I believe that the next few decades should be particularly fertile ones for statistical innovation.

Sometimes (not very often), the planets align for some lucky discipline, which then blossoms with new ideas and breathtaking progress. Microbiology is a perfect current example. The key there was a buildup of interesting questions concerning cellular processes, followed by new technology that enabled a much closer look at those processes in action.

Now the planets may be aligning for statistics. New technology—electronic computation—has broken the bottleneck of calculation that limited classical statistical theory. At the same time an onrush of important new questions has come upon us in the form of huge datasets and large-scale inference problems. I believe that the statisticians of this generation will participate in a new age of statistical innovation that might rival the golden age of Fisher, Neyman, Hotelling, and Wald.

Finally, let me thank the Association for the opportunity to serve as president, to speak here this evening, and to help honor our many deserving colleagues.

[Received January 2005. Revised January 2005.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Efron, B. (2003), "Robbins, Empirical Bayes, and Microarrays," *The Annals of Statistics*, 24, 366–378.

——— (2004). "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

Efron, B., and Tibshirani, R. (1998), "The Problem of Regions," *The Annals of Statistics*, 26, 1687–1718.

Hedenfalk, I., Duggen, D., Chen, Y. et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, 344, 539–548.

Schwartzman, A., Dougherty, R., and Taylor, J. (2005), "Cross-Subject Comparison of Principal Diffusion Direction Maps," *Magnetic Resonance in Medicine*, to appear.