THE FUTURE OF STATISTICS

Bradley Efron

The 1990 sesquicentennial celebration of the ASA included a set of eight articles on the future of statistics, published in the May issue of the American Statistician. Not all of the savants took their prediction duties seriously, but looking back from 2007 I would assign a respectable major league batting average of .333 to those who did. Of course it's only been 17 years and their average could still improve. The statistical world seems to be evolving even more rapidly today than in 1990, so I'd be wise not to sneeze at .333 as I begin my own prediction efforts.

There are at least two futures to consider: statistics as a profession, both academic and industrial, and statistics as an intellectual discipline. (And maybe a third -- demographic – with the center of mass of our field moving toward Asia, and perhaps toward female, too.) We tend to think of statistics as a small profession, but a century of steady growth, accelerating in the past twenty years, now makes "middling" a better descriptor. Together, statistics and biostatistics departments graduated some 600 Ph.D.s last year, with no shortage of good jobs awaiting them in industry or academia. The current money-making powerhouses of industry, the hedge funds, have become probability/statistics preserves. Perhaps we can hope for a phalanx of new statistics billionaires, gratefully recalling their early years in the humble ranks of the ASA.

Statistics departments have gotten bigger and more numerous around the U.S., with many universities now having two. That could become three as fields like astrostatistics and geostatistics continue to develop. One could put together another good statistics department at Stanford from the faculty in economics, business, psychology, engineering, and the medical school whose work is primarily statistical. This is the information age, statistics is the prime information science, and there is every reason to believe in a greatly increased statistical presence in the academy of the future. Or maybe not. Ideas are the coin of the realm in the intellectual world. Our continued growth and influence depends on the same thing that powered the last century, the continued production of useful new ideas and techniques.

The history of statistics in the Twentieth Century is the surprising and wonderful story of a ragtag collection of numerical methods coalescing into a central vehicle for scientific discovery. It is really two histories. The first 50 years saw the development of fundamental theory: Fisher's information theoretic approach to maximum likelihood concerned optimal estimation; then the Neyman-Pearson lemma gave us optimal testing. Together, the Fisher-Neyman-Pearson revolution raised statistics to mathematical maturity, providing statisticians the tools to attack new problems on a principled basis rather than as one-off heuristics. The resurgence of heuristics in our recent literature has something to say where the field is trying to go -- a point I'll return to in a little while.

The second half of the Twentieth Century was the methodology age. Building on our optimality foundation, and on the introduction of electronic computation, the useful scope of statistical methodology was vastly extended, removing the restriction to small normal-theory problems. I like to list in pairs a dozen post-war methodological advances: nonparametric and robust techniques, Kaplan-Meier and proportional hazards, logistic regression and general linear

models, jackknife and bootstrap methods, the EM algorithm and Gibbs sampling, and empirical Bayes and Stein estimation.

My own life as a consulting biostatistician has been revolutionized in the space of a single career. Sitting in front of a powerful terminal, calling up **R** functions to do all the dozen advances and a lot more, really does put seven-league boots on a statistician's feet.

Maybe seventy-seven league boots lie in our immediate future. I hope so: we're going to need them for the new kinds of problems scientists are bringing us these days. Technology is destiny in science, and statisticians are feeling the brunt of a technological revolution in data acquisition. The lesson of microarrays has not been lost on other scientists: modern equipment permits a thousand-fold increase in information collection. That information winds up on our desks! Statistical theory and methods are evolving in response, with the future success of the statistical enterprise heavily contingent on the quality of that response.

Articles like this one tend to be a little dreamy, because, of course, the author can't really decipher the future. As an antidote to vagueness I wanted to discuss a particular, reasonably typical, large-scale data problem, which emphasizes some of the limitations of classical theory -- and perhaps points toward what is needed in the future. The figure below concerns a study of 3747 California high schools. Standardized English exams were administered at each school, and a test statistic "**Z**" calculated, comparing economically advantaged versus disadvantaged students:

Z = {p(adv) - p(dis)} / sterr

Here **p(adv)** is the proportion of advantaged students passing the exam, similarly **p(dis)** for the disadvantaged, while **sterr** is the usual binomial estimate of standard error for the numerator.

[FIGURE ABOUT HERE]

In the first of the 3747 schools, **Z** equaled 2.28. Comparing **Z** with a standard **N(0,1)** null distribution, which is what we would almost certainly do if we had only that school's data to consider, shows advantaged students performing significantly better than disadvantaged students, two-sided **p**-value .02. This is probably true, but the figure implies that it isn't very interesting. In fact **Z = 2.28** is in the bottom quartile of the **Z** distribution -- advantaged students almost always do better in this study.

The lesson of the figure is that the totality of **Z** scores has something interesting to say about any one of them. The classic **N(0,1)** null hypothesis, which would be fine for analyzing a single school's data, seems irrelevant for all of them considered together. A normal fit to the center of the histogram suggests **N(3.38,1.77^2)** as being more appropriate here (what I like to call the "empirical null"), at least if we are looking for unusual school performances. Using the empirical null in a False Discovery Rate analysis -- another one of those useful new statistical ideas -- yielded 100 interesting schools, 95 in which the advantaged students were performing significantly better, and 5 favoring the disadvantaged. [We badly need a better word than "significant", never ideal to begin with, for those cases found interesting by a large-scale testing procedure. "Interesting" isn't technical enough, "non-null" is jargon -- maybe "detectable" or "discoverable"?]

At this point I feel the definite need for some new Fishers and Neymans. Classical theory will do a beautiful job of estimation or testing for any particular school by itself, perhaps even "adjusting for nuisance parameters" (a telling phrase), but falls silent when it comes to incorporating relevant information from the other schools, the "nuisances". So, for example, we need a new theory of information that says how well one can assess the situation in School 1 given all the data in the figure.

Committed Bayesians will feel that such a theory already exists, and has since 1763. An excellent point. Bayesian statistics excels at amassing information from disparate sources. The trick, it seems to me, is to combine Bayesian flexibility with frequentist caution -- the caution that protects a statistician from depending too much on his or her own models. Empirical Bayes methods are my favorite compromise for situations like the schools problem but we still seem to be in the Neolithic era of EB development. A relatively safe prediction is that some sort of Bayesian-frequentist compromise will blossom in the near future of statistical theory.

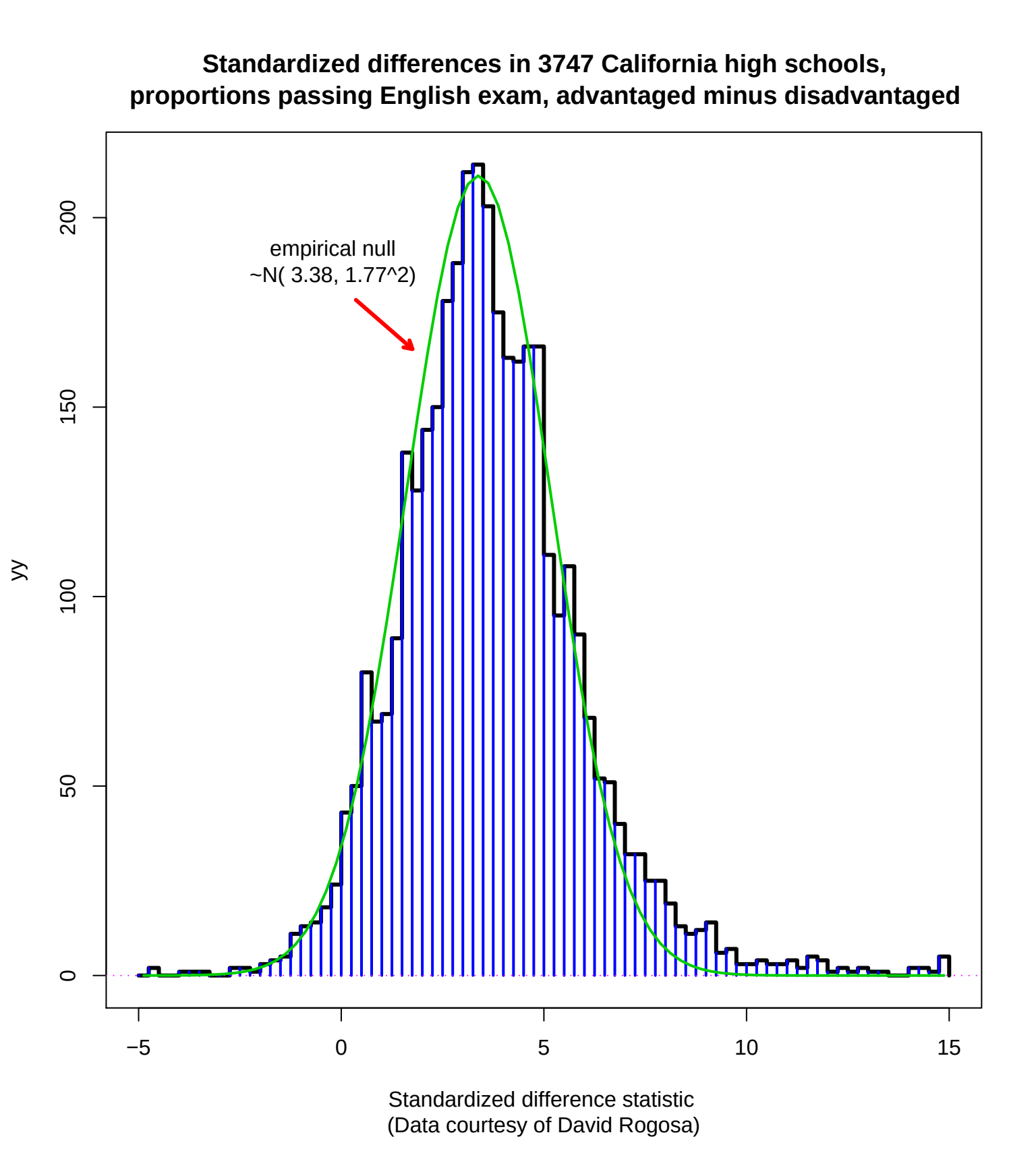
Other compromises could be in the works. Estimation and testing, the two great pillars of applied statistics, are currently pursued in quite different spirits: hypothesis testing tends toward the highly precise, with great emphasis on exact inference, ".05" having attained almost mythic status; estimation is more relaxed, allowing for example the use of regression models that are understood to be imperfect but still helpful for combining information. The schools data set is a blend of testing and estimation situations and would benefit from a combined methodology. Or perhaps I should say from a combined attitude, one that aimed for reasonably precise inferences while not doing so at the expense of wasted information.

In fact, attitude change is already in the air, at least in the world of statistical applications. A suite of problems associated with large-scale data sets -- prediction, variable selection, model building -- are under energetic development via heuristic computer-based algorithms. Often such work is neither Bayesian nor frequentist, but rather statistically agnostic, in the sense of machine learning or neural networks.

All of this reflects the limitations of classical statistics, which avoided model selection problems, "errors of the third kind", because of mathematical intractability in pre-computer times. History seems to be repeating itself: we've returned to an era of ragtag heuristics, propelled with energy but with no guiding direction. Maybe we can hope that history really will repeat itself and that some brand-new Fishers and Neymans will succeed in rationalizing all this activity over a solid theoretical foundation. Then we'll be on our way to another splendid century.

Too much to hope for? It would have seemed that way in 1900 too, but here we are. Without laying down betting money, I wouldn't be surprised to see a new golden age of statistical theory emerging in the near future.

Does all of this have anything to say to the young person about to begin a career in statistics? Well ... maybe. Statistics is in a period of rapid expansion and rapid change. During such times it pays to concentrate on basics and not tie oneself too closely to any one technology or analysis fad. I believe that you -- that hypothetical young statistician -- have an excellent shot at being in the intellectual growth industry of the Twenty-First Century.


Standardized differences in 3747 California high schools,
proportions passing English exam, advantaged minus disadvantaged
empirical null
~N(3.38, 1.77^2)
yy
0
50
100
150
200
−5
0
5
10
15
Standardized difference statistic
(Data courtesy of David Rogosa)