The Future of Statistics

Bradley Efron
Stanford University

"Strange, as one gets older you're expected to know more about the future."

The history of statistics as a recognized discipline divides rather neatly at 1900, the year of Karl Pearson's chi-square paper. Before then we are still close to the world of Quetelet, where huge census-level data sets are brought to bear on simple but important questions: Are there more male or female births? Is the murder rate rising? Then, as if on cue, the Twentieth Century rings in a focus on small-scale statistics. A team of intellectural giants, Fisher, Neyman, Hotelling, ..., invent a theory of optimal inference, capable of wringing out every drop of collected information. The questions are still simple: Is treatment A better than treatment B? But the new methods are suited to the kinds of small data sets an individual scientist might collect.

What does this have to do with the future of statistics? Quite a bit, perhaps: the Twenty-First Century, again on cue, seems to have initiated a third statistical era. New technologies, exemplified by the microarray, permit scientists to collect their own huge data sets. But this is not a return to the age of Quetelet. The flood of data is now accompanied by a flood of questions, perhaps thousands of them, that the statistician is charged with answering together; not at all the setting Fisher et al. had in mind.

As a cruder summary of my already crude statistical history, we have

19th Century:	Large data sets, simple questions
20th Century:	Small data sets, simple questions
21st Century:	Large data sets, complex questions

The future of statistics, or at least the next large chunk of future, will be preoccupied, I believe, with problems of large-scale inference raised in our revolutionary scientific environment. For example, how should one analyze $10,000$ related hypothesis tests or $100,000$ correlated estimates at the same time?

Figure 1 concerns an example of large-scale inference from Singh et al. (2002): 52 prostate cancer patients and 50 normal controls have each had his genetic expression levels measured on $N = 6033$ genes. This produces a matrix of measurements $\boldsymbol{X}$ with $N = 6033$ rows, one for each gene, and 102 columns, one for each man: enormous by Twentieth Century standards but nothing remarkable these days. We wonder which of the genes, if any, are more active in the cancer patients.

As a first step we can compute a two-sample t-statistic t_i comparing expression levels between cancer patients and controls on gene i. For Figure 1, each t_i has been transformed into a z-value z_i, by definition a test statistic having a standard normal distribution under the null hypothesis that gene i behaves the same in both groups,

$$H_0 : z_i \sim \mathcal{N}(0, 1). \tag{1}$$

The histogram of the 6033 z_i's looks like a $\mathcal{N}(0,1)$ curve near its center, which makes sense since presumably most of the genes are *not* involved in prostate cancer etiology, but it also shows a promising excess of values in the extreme tails. For example, 49 of the z_i's exceed 3 (indicated by the hash marks) whereas the expected number is only 8.14 if all the genes follow (1). Should we report the list of 49 back to the researchers as interesting candidates for further study?

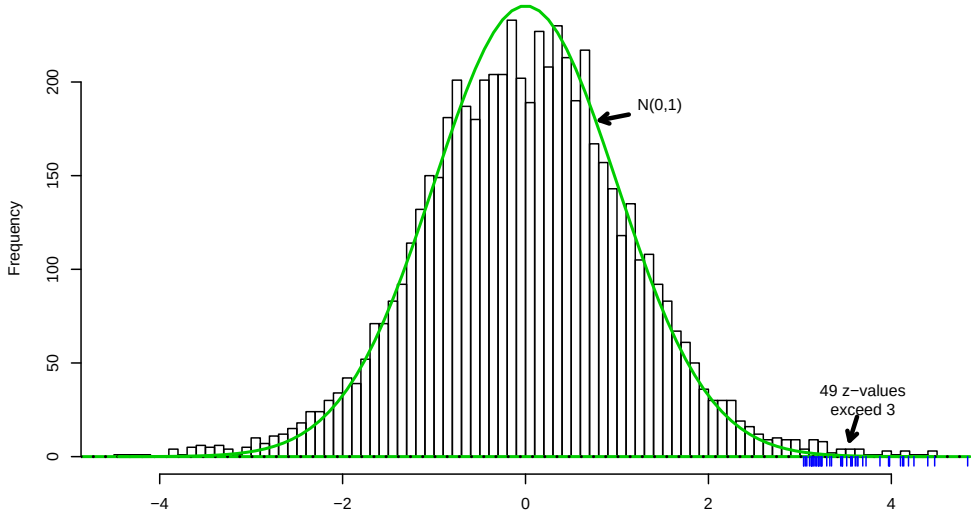

Figure 1: $N = 6033$ z-values, prostate study

Any one of the genes is wildly significant by classical single-test standards where we would reject H_0 for $z_i > 1.96$, the two-sided .05 value. But with $N = 6033$, the .05 Bonferroni bound requires $z_i > 4.46$, the two-sided $.05/N$ value, and only 3 of the 49 genes make the cut.

In what might be taken as a premonitory salvo of Twenty-First Century statistics, Benjamini and Hochberg (1995) proposed a different, more lenient standard for large-scale testing based on *False Discovery Rates*:

$$\text{Fdr}(3) = 8.14/49 = .166 \tag{2}$$

in our case, the ratio of counts expected under null conditions to those actually observed in the interval $(3, \infty)$. Assuming independence of the z-values, they showed that a statistician who chooses to reject all z_i's in the largest interval (x, ∞) such that $\text{Fdr}(x)$ is less than some control level q will make an expected proportion of false discoveries no greater than q. Taking $q = .166$ for the prostate data gives $x = 3$ and suggests that 1/6 of the list of 49 are false discoveries, the other 5/6 being genuinely non-null genes: not bad odds for the prospects of further investigation.

Controlling Fdr is fundamentally different than controlling the probability of Type I error. Now the significance of a gene that has $z_i > 3$ depends on how many others exceed 3. If there were only 10 such, instead of 49, we would have $\text{Fdr}(3) = .81$; not an encouraging prospect for the investigators.

Twentieth Century applied statistics has been very much a world of direct evidence in which each case, each gene in our example, is judged entirely on its own data. This is a world designed for frequentism, where objectivity is enforced by notions of unbiasedness, minimum variance, size and power. But large-scale data sets like that for the prostate study abound with indirect evidence: our interest in z_i is affected by all the other z_j's. I believe that the immediate future of statistical theory and practice crucially involves "learning from the experience of others", i.e., the incorporation of indirect evidence.

Bayes theorem is a perfect recipe for learning from the experience of others, and we can expect Bayesian methods to play a greater role in Twenty-First Century data analysis. Fdr theory was derived frequentistically, but it has a compelling Bayesian rationale. Assuming that the prior probability of a null case is near 1, Bayes theorem yields

$$\Pr\{\text{gene } i \text{ is null} | z_i \geq x\} \doteq F_0(x)/F(x) \tag{3}$$

where F_0 is the probability that a null z_i exceeds x [equaling $1 - \Phi(x)$ under (1)] and $F(x)$ is the probability that a randomly selected z_i, null or not, exceeds x. Substituting the empirical cdf $\hat{F}(x)$ for the unknown $F(x)$ gets us back to definition (2); see Efron (2008). We can restate the Benjamini–Hochberg procedure in Bayesian terms: "Reject those z_i's in the largest interval (x, ∞) that has estimated Bayes null probability (3) less than q."

Indirect evidence is not the sole property of Bayesians. Tukey's phrase "borrowing strength" nicely captures the frequentist regression tactic of using nearby data points to assist in estimation at a particular point of interest. "Nearby" refers to distance in a space of relevant covariates. The explosion in data collection has brought with it an explosion in the number of covariates, often too many for standard regression techniques. A thriving industry of new methods has emerged — boosting, bagging, CART, Lasso, LARS, projection pursuit — which search to build effective regression models from subsets of the available regressors. The generic term here is *data mining*, which began as an insult but now seems to have its own robust statistical future.

Bayesian and frequentist ideas are combined happily in the Fdr algorithm. Other lines are blurred too: in (2) we are *estimating* the *hypothesis testing* quantity (3); that is, we are carrying out an "empirical Bayes" analysis, to use Robbins' apt description. Blurred lines are another likely (and hopeful) trend, as Twenty-First Century statisticians outgrow the confines of classical theory.

In moving beyond the classical confines we are also moving outside its wall of protection. Fisher, Neyman et al. fashioned, with enormous intellectual effort, an almost perfect inferential machine for small-scale estimation and testing problems. It took our brilliant predecessors at least 25 years to work the kinks out of ANOVA/linear model theory. My guess is for another long period of progress and retrenchment. Difficulties with large-scale inference are easy to find. Not all microarray data sets are as obliging as that from the prostate study. Often the histogram is much wider or narrower than in Figure 1, casting grave doubt on the adequacy of the textbook null hypothesis (1). Correlations sprawl across the z-values, undermining accuracy of the empirical Bayes estimator. Effect sizes for the genes deemed "non-null" are difficult to assess because of massive selection biases from choosing among so many candidates. Et cetera, et cetera. In other words, there is a lot for statisticians to think about over the next 25 years.

Some of that thinking will involve the legitimate use of Bayesian methods in large-scale inference. As an example, Figure 2 concerns effect size estimation in the prostate study. Suppose that the z-value for gene i follows a normal distribution,

$$z_i \sim \mathcal{N}(\mu_i, 1), \tag{4}$$

μ_i the "effect size" (so $\mu_i = 0$ for the null genes (1)), where the effect sizes are drawn from some unknown Bayesian prior distribution $G(\mu)$. Call $f(z)$ the marginal density of z-values induced by $G(\mu)$ and (4). Then it is not difficult to show that expected effect size $\mu(z)$ is a simple function of $f(z)$,

$$\mu(z) = z + \frac{d}{dz} \log f(z). \tag{5}$$

The heavy curve in Figure 2 is an empirical Bayes estimate of (5): a smooth curve $\hat{f}(z)$ was fit to the heights of the histogram bars in Figure 1 and its logarithm differentiated to give $\hat{\mu}(z)$; see

Efron (2009). Gene 610 has $z_{610} = 5.29$, the largest of the 6033 z-values, with effect size estimate $\hat{\mu}_{610} = 4.11$, as indicated.

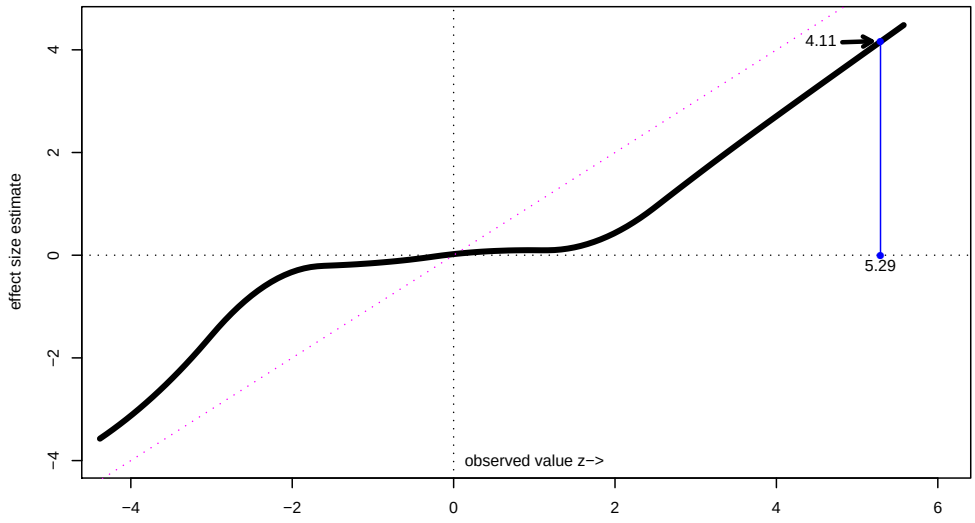

Figure 2: Effect size estimates for prostate study

We can be almost certain that z_{610}, as the maximum of $N = 6033$ observations, exaggerates μ_{610}. This is the curse of selection bias. Bayes estimates, according to theory, are immune to selection bias: if $\hat{\mu}_{610}$ were a genuine Bayes posterior estimate of μ_{610}, we would not have to concern ourselves that gene 610 was selected on the basis of the data. A Bayes prior effectively postulates an *infinite* amount of relevant past experience, swamping selection effects from mere thousands of observations.

Are empirical Bayes estimates immune to selection bias? $N = 6033$ is not $N =$ infinity, and we might suspect at least some selection effects to linger in $\hat{\mu}_{610}$. Other very appealing Bayesian properties, like the right to take interim looks at a clinical trial without an optional stopping penalty, are not put to the test in small-scale situations. Large-scale studies offer their own self-contained universes in which it may be possible to settle such questions.

Statistics is a second-level science: our "nature" is the data-analytic questions posed by front-line scientists — biologists, astronomers, economists, etc. — the "etc." by now spanning almost all areas of quantitative inquiry. The future of statistics depends on the future of science, in particular of scientific technology. There's a good chance that today's huge data sets will seem puny in a few years, in which case this little essay will look remarkably timid.

The future I've been discussing is that of statistics as an intellectual discipline. What about the future of the statistics profession? There is no question that the probabilistic/statistical point of view continues its relentless spread across science and engineering. (Maybe scientists are just running out of problems simple enough to solve deterministically.) So there will be more people interested in statistical questions, but that doesn't necessarily imply more statisticians. The field could fractionate into subject area subspecialties.

I don't think so. The health of a scientific profession can be rated on three criteria:

- An outside demand for answers in the profession's chosen area.
- Some evidence of past success in asnwering such questions.
- An ongoing production of useful new ideas.

In other words, the profession should be healthy from both an inside and outside point of view. I give statistics high grades on all three criteria, perhaps higher than at any time in the past half-century. In one sense we have a monopoly: statistics is the only profession that takes applied inference seriously as a subject of study. So in my view, our future work is cut out for us, but it's not cut out for anyone else.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57: 289–300.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* 23: 1–22.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* To appear.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203–209.