Statistical Thinking for 21st Century Scientists

D.R. Cox
Nuffield College, Oxford

Bradley Efron
Stanford University

Statistical science provides a wide range of concepts and methods for studying situations subject to unexplained variability. Such considerations enter fields ranging from particle physics and astrophysics through to genetics and sociology and economics, and even beyond, and, further, to the associated areas of application like engineering, agriculture and medicine, in the last especially but by no means only in clinical trials. Successful application hinges on absorption of statistical thinking into the subject matter and as such depends strongly on the field in question and on the individual investigators. It is the job of theoretical statisticians both to be alive to the challenges of specific applications and at the same time to develop methods and concepts which, with good fortune, will be broadly applicable.

To illustrate the breadth of statistical concepts it is helpful to think of the following sequence, in practice often encountered in a different order:

1. Clarification of research questions in a complex situation.

2. Specification of the context for study, for example the choice of individuals for entry into a clinical trial.

3. Issues of metrology: How are key features best measured in the context in question and how secure is the measurement process?

Considered broadly, there may be many aspects of study design. General aims are to achieve a reasonable level of precision, an absence of systematic error and economy and breadth of interpretation, sometimes by answering several interconnected questions in one study.

- Data collection, possibly including monitoring of data quality.

- Data analysis, usually in various stages from the simple descriptive onwards.

- Summary of conclusions.

- Interpretation: What is the underlying interpretation of what has been found? What are the relations with other work in the field? What new questions have been raised?

General statistical considerations may enter at all these stages even though in essence they are all key subject matter concerns.

Phrases often heard nowadays are big data, machine learning, data science, and, most recently, deep learning. Big data have been around a long time but what, of course, is new is

the ability to analyse such data other than on a sampling basis. Key issues concern first the relevance of the data, especially if it is collected in a sense fortuitously. Then there may or should be worries over quality. Some big data, for example that obtained in the investigation at CERN leading to the Higgs boson, is of very high quality. In other situations, however, if a small amount of bad data may be quite misleading a very large amount of bad data may be exceedingly misleading. The third aspect is more technically statistical. The simpler methods of precision assessment may appear to indicate a very narrow confidence band on the conclusions from big data and this narrowness may give a seriously overoptimistic view of the precision achieved.

The other newer themes involve important ideas coming with heavy computer science emphasis and often aimed at empirical prediction from noisy data rather than either with probing the underlying interpretation of the data or with issues of study design or with the nature of the measurement process.

The theory and practice of computer-age statistics is, for the most part, a case of new wine in old bottles: the fundamental tenets of good statistical thinking haven't changed, but their implementation has. This has been a matter of necessity. Data collection for a modern scientist can move in seven-league boots thanks to spectacular advancements in equipment – notable examples include microarrays and DNA sequencers in microbiology, and robotic telemetry for astronomy. Along with Big Data comes Big Questions, often thousands of hypothesis testing and estimation problems posed simultaneously demanding careful statistical discussion.

Statisticians have responded with much more flexible and capacious analysis methods. These depend of course on the might of modern computation, but also on powerful extensions of classical theories, that shift the burden of mathematical analysis onto computable algorithms – but demand careful discussion for the formulation of principles. The examples which follow are too small to qualify as Big Data, but, hopefully, are big enough to get the idea across.

A study at a pediatric hospital in Guatemala followed some 1800 children over a 12-year period beginning in 2002 [1]. Ten percent of the children were abandoned by their families during their stay. The goal of the study was to identify the causes of abandonment. The key response variable was **Time**, the number of days from admission to abandonment. For 90% of the children, abandonment was never observed, due to leaving the hospital or the study period ending, in which case **Time** was known only to *exceed* the number of days of observation. In common terminology, **Time** was heavily censored.

More than forty possible explanatory factors were measured, only six of which will be discussed here: **Distance**, the distance of the child's home from the hospital; **Date**, the date of the childs admission measured in days since the study's beginning; **Age** and **Sex** of the child; and **ALL** or **AML**, indicating that the child was suffering from acute lymphoblastic leukemia or acute myeloid leukemia (a worse prognosis). All of the variables were standardized. (Note: ALL and AML are two of several diagnoses under consideration, all of which were considered "others" for this analysis.)

Proportional hazards is a modern regression methodology that allows the fair comparison of potentially causative factors for a censored response variable [2]. Table 1 shows its output for the abandonment study. **Date** for instance has a very strongly negative estimate, indicating that abandonment was decreasing as calendar time went on. **Distance** was strongly

Table 1: Proportional hazards analysis of the abandonment data. Estimated **Date** coefficient 1.660 is strongly negative, indicating decreased abandonment as study progressed.

	Estimate	Standard Error	z-Value	p-Value	Bootstrap Standard Error
Distance	0.210	0.072	2.902	0.004	0.068
Date	−1.660	0.107	−15.508	0.000	0.088
Age	−0.154	0.084	−1.834	0.067	0.082
Sex	−0.027	0.076	−0.347	0.729	0.078
ALL	0.146	0.082	1.771	0.077	0.083
AML	−0.070	0.081	−0.864	0.387	0.088

positive, suggesting increased abandonment from remote home locations. Neither **Age** nor **Sex** yielded significant p-values, though there is some suggestion that older children did better. Likewise, neither **ALL** or **AML** achieved significance but, perhaps surprisingly, AML children seemed better off.

In addition to the parameter estimates, first column of Table 1, proportional hazards theory also provides approximate standard errors, column 2. The bootstrap [3] was used as a check. Each bootstrap data set was formed by sampling the 1800 children 1800 times *with replacement*; so child 1 might appear twice, child 2 not at all, child 3 once, etc. Then the proportional hazards model was run for the bootstrap data set, giving new estimates for **Distance**, **Date**, **Age**, **Sex**, **ALL**, and **AML**. Two thousand bootstrap data sets were independently generated, yielding the bootstrap standard errors in column 5 of the table. For instance, the 2000 bootstrap estimates for **Distance** had empirical standard deviation 0.068, nearly the same as the theoretical standard error 0.072. With the moderate exception of **Date**, the other comparisons were similarly reassuring.

The bootstrap replications can be used to address a variety of other inferential questions. Figure 1 shows the histogram of the 2000 bootstrap estimates of the difference **AML** minus **ALL**. Only 34 of the 2000 exceed zero, yielding a one-sided bootstrap p-value of 0.017 ($= 34/2000$) against the null hypothesis of no difference.

The proportional hazards algorithm required perhaps 100 times as much computation as a standard linear regression, while the bootstrap analysis multiplied the burden by 2000. Neither theory would have been formulated in the age of mechanical calculation. They are discussed in Chapters 9 and 10 of [4], along with a suite of other computer-intensive statistical inference methods.

At a fundamental level, statistical theory concerns learning from experience, especially from experience that arrives a little bit at a time, perhaps in noisy and partly contradictory forms. Modern equipment allows modern scientists to cast wider experiential nets. This has increased the burden on the statistical learning portion of the scientific process. Our next example, taken from Chapter 6 of [4], shows the learning process in action, using statistical ideas proposed in the 1950s, but only now routinely feasible.

Figure 2 concerns a study of 844 patients undergoing surgery for stomach cancer. Besides the removal of the central site, surgeons often remove surrounding lymph nodes, which are

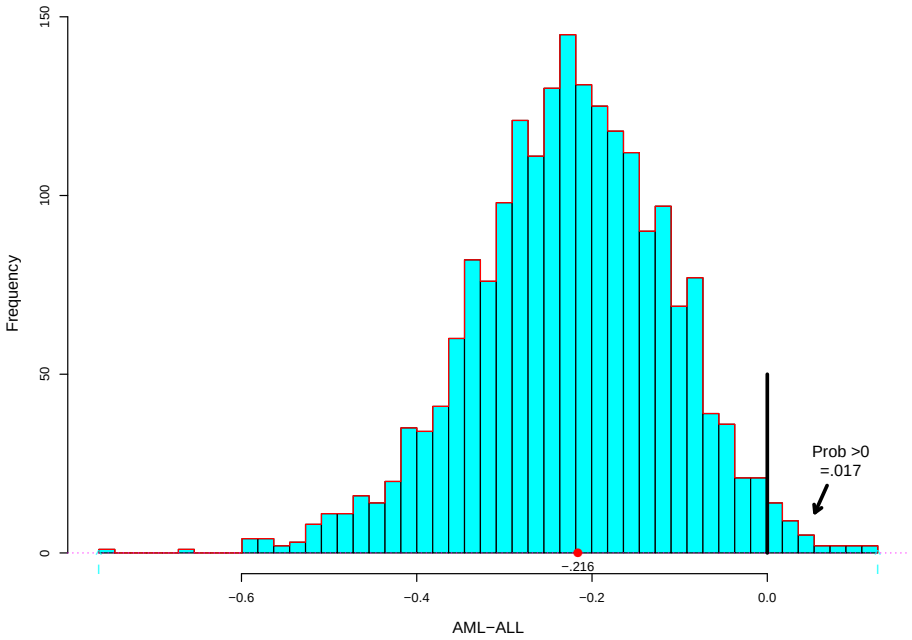

Figure 1: 2000 bootstrap replications of difference between AML and ALL proportional hazards coefficients.

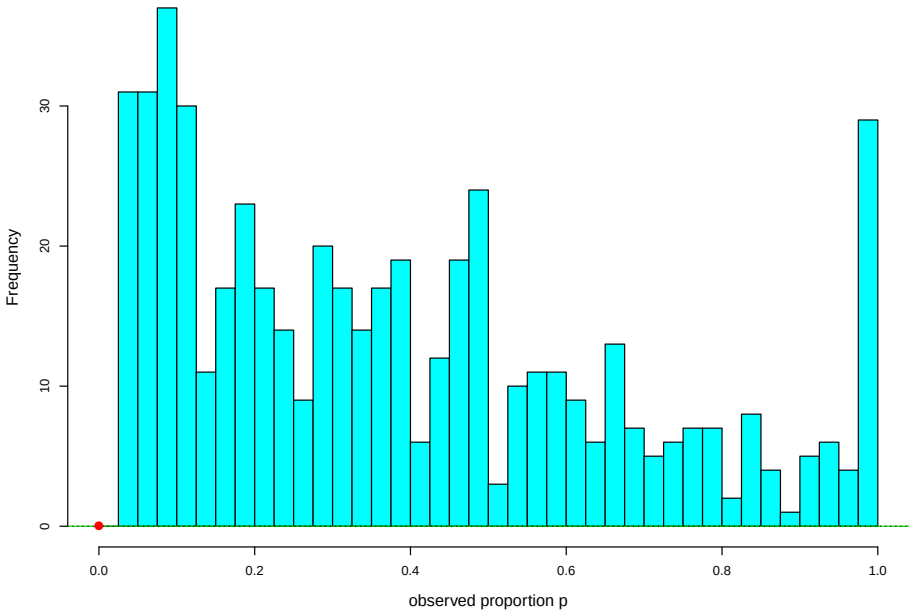

Figure 2: Observed proportion p of malignant nodes for 522 patients having $p > 0$; 322 patients (38%) had $p = 0$, indicated by the large dot.

subsequently evaluated as positive (malignant) or negative. For patient i, $i = 1, 2, \ldots, 844$, let

$$n_i = \#\text{ nodes removed}, \qquad x_i = \#\text{ nodes positive},$$

and

$$p_i = x_i/n_i,$$

the proportion of positive nodes; n_i varied between 1 and 69. The histogram in Figure 2 depicts the 522 patients with $p_i > 0$, i.e., having at least one positive node; 322 of the patients, about 38%, had $p_i = 0$, represented by the large dot.

It is reasonable to imagine that each patient has a *frailty parameter* θ_i, indicating how prone he or she is to positive nodes, and that we are seeing binomial observations

$$x_i \sim \text{Binomial}(n_i, \theta_i);$$

equivalently, x_i is the number of heads observed in n_i independent flips of a coin having probability of heads θ_i. If the n_i's were very large then $p_i = x_i/n_i$ would nearly equal θ_i. However, many of the n_i were quite small (eight of them equaling 1) so in fact Figure 2 gives a badly distorted picture of the distribution of the θ_i's.

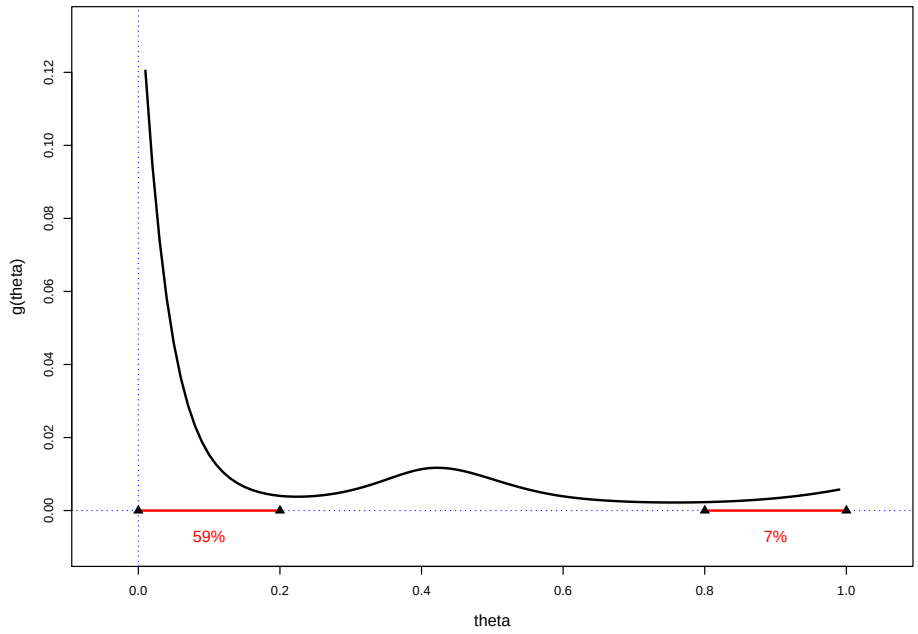

Figure 3: Estimated prior density for frailty parameter θ; median value $\theta = 0.09$.

Empirical Bayes methods allow us to recover a good estimate of what a histogram of the 844 true θ_i values would look like. We assume that the θ_i's have some prior density $g(\theta)$; $g(\theta)$ is unknown but assumed to belong to a low-dimensional parametric family. Here $\log g(\theta)$ was assumed to be a fifth-order polynomial in θ. Maximizing the likelihood of the observed data (n_i, x_i), $i = 1, 2, \ldots, 844$, over the coefficients of the polynomial yielded the

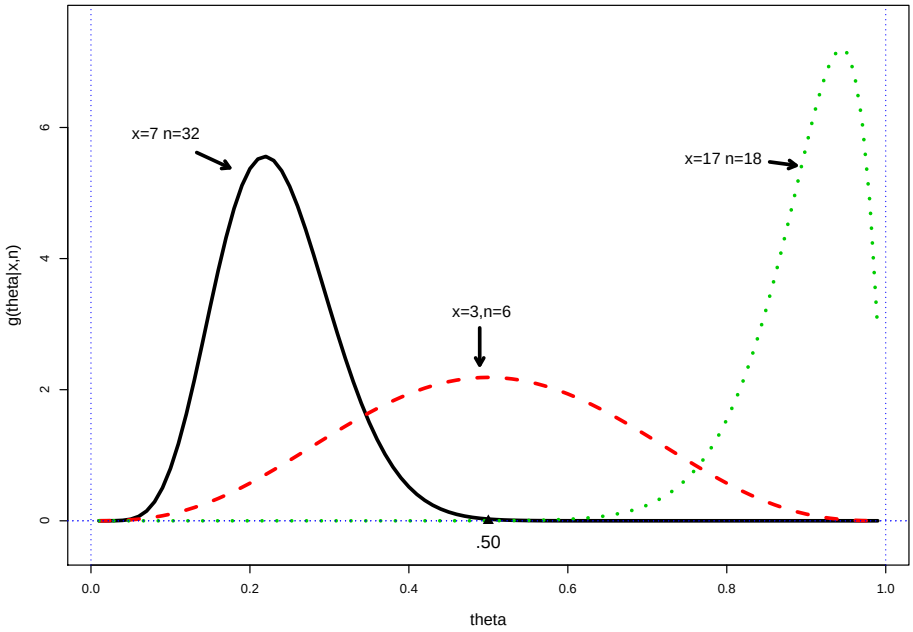

Figure 4: Posterior probabilities of frailty parameter θ for three hypothetical patients.

estimate of $g(\theta)$ pictured in Figure 3. It shows that most of the frailties are small, 59% less than 0.2, but there are large ones too, 7% above 0.8.

Having estimated the prior density $g(\theta)$, we can employ Bayes rule to calculate the posterior density of θ_i given patient i's observed values n_i and x_i. This is done for three of the patients in Figure 4. Patient 1, with $n_i = 32$ and $x_i = 7$, is seen to almost certainly have θ_i less than 0.5; Patient 3, with $n_i = 18$ and $x_i = 17$, almost certainly has frailty θ_i greater than 0.5; Patient 2, with $n_i = 6$ and $x_i = 3$, could conceivably have almost any value of θ_i. This kind of information may be valuable for recommending follow-up therapy that is either more stringent or less.

The observed data $n_1 = 32$ and $x_1 = 7$ represents *direct* statistical evidence for Patient 1. It provides, among other things, the direct estimate $p_1 = 7/32 = 0.22$ for θ_1. *Indirect* evidence, from the other 843 patients, also contributed to the posterior probability density for Patient 1 seen in Figure 4.

An increased acceptance of indirect evidence is a hallmark of modern statistical practice. Both frequentist techniques (regression algorithms) and Bayesian methods are combined in an effort to bring enormous amounts of possibly relevant "other" cases to bear on a single case of particular interest, i.e., Patient 1 in the nodes study. Avoiding the difficulties and pitfalls of indirect evidence motivates much of current statistical research.

We emphasize the current high level of fruitful application and methodological development. This is, however, anchored in a long history going back in particular to the great early 19th century mathematicians, Gauss and Laplace. Their statistical work was motivated by concerns over the analysis of astronomical data. Quasi-philosophical disagreements over the meaning of probability have rumbled on since then. Our attitude is eclectic, but in the last analysis we see a contrast, not a conflict, between the use of probability to represent in idealized form patterns of variability in the real world and its use to capture the uncertainty of

our conclusions. Controversy centres mostly on the second and more than one approach may be fruitful. In the last analysis, however, we are using probability as a measuring instrument and in some sense it must be well calibrated.

We have worked as statisticians for a combined total of 125 years $(72 + 53)$ and both of us fully retain our enthusiasm for the field. It has changed enormously over our lifetimes and no doubt will continue to do so. But at the heart of our subject are core issues about uncertainty and variability that have both a permanent value and an exciting continuing challenge that is conceptual, mathematical and computational.

References

[1] E. Alvarez, M. Seppa, K. Messacar, J. Kurap, et al. Improvement of abandonment of therapy in pediatric patients with cancer in Guatemala. *J. Global Onc.*, online July 20, 2016. doi: 10.1200/JGO.2016.004648. 4th Annual Symposium on Global Cancer Research (April 8, 2016).

[2] D. R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. ISSN 0035-9246. URL `jstor.org/stable/2985181`. With discussion and a reply by the author.

[3] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1): 1–26, 1979. ISSN 0090-5364. MR515681.

[4] Efron Bradley and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, Cambridge, 2016. ISBN 13: 978-1107149892. Institute of Mathematical Statistics Monographs (Book 5).