Statistical Science
1998, Vol. 13, No. 2, 95–122

R. A. Fisher in the 21st Century

Invited Paper Presented at the 1996 R. A. Fisher Lecture

Bradley Efron

Abstract. Fisher is the single most important figure in 20th century statistics. This talk examines his influence on modern statistical thinking, trying to predict how Fisherian we can expect the 21st century to be. Fisher's philosophy is characterized as a series of shrewd compromises between the Bayesian and frequentist viewpoints, augmented by some unique characteristics that are particularly useful in applied problems. Several current research topics are examined with an eye toward Fisherian influence, or the lack of it, and what this portends for future statistical developments. Based on the 1996 Fisher lecture, the article closely follows the text of that talk.

Key words and phrases: Statistical inference, Bayes, frequentist, fiducial, empirical Bayes, model selection, bootstrap, confidence intervals.

1. INTRODUCTION

Even scientists need their heroes, and R. A. Fisher was certainly the hero of 20th century statistics. His ideas dominated and transformed our field to an extent a Caesar or an Alexander might have envied. Most of this happened in the second quarter of the century, but by the time of my own education Fisher had been reduced to a somewhat minor figure in American academic statistics, with the influence of Neyman and Wald rising to their high water mark.

There has been a late 20th century resurgence of interest in Fisherian statistics, in England where his influence never much waned, but also in America and the rest of the statistical world. Much of this revival has gone unnoticed because it is hidden behind the dazzle of modern computational methods. One of my main goals here will be to clarify Fisher's influence on modern statistics. Both the strengths and limitations of Fisherian thinking will be described, mainly by example, finally leading up to some speculations on Fisher's role in the statistical world of the 21st century.

What follows is basically the text of the Fisher lecture presented to the August 1966 Joint Statistical meetings in Chicago. The talk format has certain advantages over a standard journal article. First and foremost, it is meant to be absorbed quickly, in an hour, forcing the presentation to concentrate on main points rather than technical details. Spoken language tends to be livelier than the gray prose of a journal paper. A talk encourages bolder distinctions and personal opinions, which are dangerously vulnerable in a written article but appropriate I believe for speculations about the future. In other words, this will be a broad-brush painting, long on color but short on detail.

These advantages may be viewed in a less favorable light by the careful reader. Fisher's mathematical arguments are beautiful in their power and economy, and most of that is missing here. The broad brush strokes sometimes conceal important areas of controversy. Most of the argumentation is by example rather than theory, with examples from my own work playing an exaggerated role. References are minimal, and not indicated in the usual author–year format but rather collected in annotated form at the end of the text. Most seriously, the one-hour limit required a somewhat arbitrary selection of topics, and in doing so I concentrated on

Bradley Efron is Max H. Stein Professor of Humanities and Sciences and Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, California 94305-4065 (e-mail: brad@stat.stanford.edu).

those parts of Fisher's work that have been most important to me, omitting whole areas of Fisherian influence such as randomization and experimental design. The result is more a personal essay than a systematic survey.

This is a talk (as I will now refer to it) on Fisher's influence, not mainly on Fisher himself or even his intellectual history. A much more thorough study of the work itself appears in L. J. Savage's famous talk and essay, "On rereading R. A. Fisher," the 1971 Fisher lecture, a brilliant account of Fisher's statistical ideas as sympathetically viewed by a leading Bayesian (Savage, 1976). Thanks to John Pratt's editorial efforts, Savage's talk appeared, posthumously, in the 1976 *Annals of Statistics*. In the article's discussion, Oscar Kempthorne called it the best statistics talk he had ever heard, and Churchill Eisenhart said the same. Another fine reference is Yates and Mather's introduction to the 1971 five-volume set of Fisher's collected works. The definitive Fisher reference in Joan Fisher Box's 1978 biography, *The Life of a Scientist*.

It is a good rule never to meet your heroes. I inadvertently followed this rule when Fisher spoke at the Stanford Medical School in 1961, without notice to the Statistics Department. The strength of Fisher's powerful personality is missing from this talk, but not I hope the strength of his ideas. Heroic is a good word for Fisher's attempts to change statistical thinking, attempts that had a profound influence on this century's development of statistics into a major force on the scientific landscape. "What about the next century?" is the implicit question asked in the title, but I won't try to address that question until later.

2. THE STATISTICAL CENTURY

Despite its title, the greater portion of the talk concerns the past and the present. I am going to begin by looking back on statistics in the 20th century, which has been a time of great advancement for our profession. During the 20th century statistical thinking and methodology have become the scientific framework for literally dozens of fields, including education, agriculture, economics, biology and medicine, and with increasing influence recently on the hard sciences such as astronomy, geology and physics.

In other words, we have grown from a small obscure field into a big obscure field. Most people and even most scientists still don't know much about statistics except that there is something good about the number ".05" and perhaps something bad about the bell curve. But I believe that this will change in the 21st century and that statistical methods will be widely recognized as a central element of scientific thinking.

The 20th century began on an auspicious statistical note with the appearance of Karl Pearson's famous χ^2 paper in the spring of 1900. The groundwork for statistics's growth was laid by a pre–World War II collection of intellectual giants: Neyman, the Pearsons, Student, Kolmogorov, Hotelling and Wald, with Neyman's work being especially influential. But from our viewpoint at the century's end, or at least from my viewpoint, the dominant figure has been R. A. Fisher. Fisher's influence is especially pervasive in statistical applications, but it also runs through the pages of our theoretical journals. With the end of the century in view this seemed like a good occasion for taking stock of the vitality of Fisher's legacy and its potential for future development.

A more accurate but less provocative title for this talk would have been "Fisher's influence on modern statistics." What I will mostly do is examine some topics of current interest and assess how much Fisher's ideas have or have not influenced them. The central part of the talk concerns six research areas of current interest that I think will be important during the next couple of decades. This will also give me a chance to say something about the kinds of applied problems we might be dealing with soon, and whether or not Fisherian statistics is going to be of much help with them.

First though I want to give a brief review of Fisher's ideas and the ideas he was reacting to. One difficulty in assessing the importance of Fisherian statistics is that it's hard to say just what it is. Fisher had an amazing number of important ideas and some of them, like randomization inference and conditionality, are contradictory. It's a little as if in economics Marx, Adam Smith and Keynes turned out to be the same person. So I am just going to outline some of the main Fisherian themes, with no attempt at completeness or philosophical reconciliation. This and the rest of the talk will be very short on references and details, especially technical details, which I will try to avoid entirely.

In 1910, two years before the 20-year-old Fisher published his first paper, an inventory of the statistics world's great ideas would have included the following impressive list: Bayes theorem, least squares, the normal distribution and the central limit theorem, binomial and Poisson methods for count data, Galton's correlation and regression, multivariate distributions, Pearson's χ^2 and Student's t. What was missing was a core for these ideas. The list existed as an ingenious collection of ad hoc devices. The situation for statistics was similar to the one now faced by computer science.

In Joan Fisher Box's words, "The whole field was like an unexplored archaeological site, its structure hardly perceptible above the accretions of rubble, its treasures scattered throughout the literature."

There were two obvious candidates to provide a statistical core: "objective" Bayesian statistics in the Laplace tradition of using uniform priors for unknown parameters, and a rough frequentism exemplified by Pearson's χ^2 test. In fact, Pearson was working on a core program of his own through his system of Pearson distributions and the method of moments.

By 1925, Fisher had provided a central core for statistics—one that was quite different and more compelling than either the Laplacian or Pearsonian schemes. The great 1925 paper already contains most of the main elements of Fisherian estimation theory: consistency; sufficiency; likelihood; Fisher information; efficiency; and the asymptotic optimality of the maximum likelihood estimator. Partly missing is ancillarity, which is mentioned but not fully developed until the 1934 paper.

The 1925 paper even contains a fascinating and still controversial section on what Rao has called the second order efficiency of the maximum likelihood estimate (MLE). Fisher, never really satisfied with asymptotic results, says that in small samples the MLE loses less information than competing asymptotically efficient estimators, and implies that this helps solve the problem of small-sample inference (at which point Savage wonders why one should care about the amount of information in a point estimator).

Fisher's great accomplishment was to provide an optimality standard for statistical estimation—a yardstick of the best it's possible to do in any given estimation problem. Moreover, he provided a practical method, maximum likelihood, that quite reliably produces estimators coming close to the ideal optimum even in small samples.

Optimality results are a mark of scientific maturity. I mark 1925 as the year statistical theory came of age, the year statistics went from an ad hoc collection of ingenious techniques to a coherent discipline. Statistics was lucky to get a Fisher at the beginning of the 20th century. We badly need another one to begin the 21st, as will be discussed near the end of the talk.

3. THE LOGIC OF STATISTICAL INFERENCE

Fisher believed that there must exist a logic of inductive inference that would yield a correct answer to any statistical problem, in the same way that ordinary logic solves deductive problems. By using such an inductive logic the statistician would be freed from the a priori assumptions of the Bayesian school.

Fisher's main tactic was to logically reduce a given inference problem, sometimes a very complicated one, to a simple form where everyone should agree that the answer is obvious. His favorite target for the "obvious" was the situation where we observe a single normally distributed quantity x with unknown expectation θ,

$$x \sim N(\theta, \sigma^2), \tag{1}$$

the variance σ^2 being known. Everyone agrees, says Fisher, that in this case, the best estimate is $\hat{\theta} = x$ and the correct 90% confidence interval for 0 (to use terminology Fisher hated) is

$$\hat{\theta} \pm 1.645\sigma. \tag{2}$$

Fisher's inductive logic might be called a theory of types, in which problems are reduced to a small catalogue of obvious situations. This had been tried before in statistics, the Pearson system being a good example, but never so forcefully nor successfully. Fisher was astoundingly resourceful at reducing problems to simple forms like (1). Some of the devices he invented for this purpose were sufficiency, ancillarity and conditionality, transformations, pivotal methods, geometric arguments, randomization inference and asymptotic maximum likelihood theory. Only one major reduction principle has been added to this list since Fisher's time, invariance, and that one is not in universal favor these days.

Fisher always preferred exact small-sample results but the asymptotic optimality of the MLE has been by far the most influential, or at least the most popular, of his reduction principles. The 1925 paper shows that in large samples the MLE $\hat{\theta}$ of an unknown parameter θ approaches the ideal form (1),

$$\hat{\theta} \to N(\theta, \sigma^2),$$

with the variance σ^2 determined by the Fisher information and the sample size. Moreover, no other "reasonable" estimator of θ has a smaller asymptotic variance. In other words, the maximum likelihood method automatically produces an estimator that can reasonably be termed "optimal," without ever invoking the Bayes theorem.

Fisher's great accomplishment triggered a burst of interest in optimality results. The most spectacular product of this burst was the Neyman–Pearson lemma for optimal hypothesis testing, followed soon by Neyman's theory of confidence intervals. The Neyman–Pearson lemma did for hypothesis testing what Fisher's MLE theory did for estimation, by pointing the way toward optimality.

Philosophically, the Neyman–Pearson lemma fits in well with Fisher's program: using mathematical logic it reduces a complicated problem to an obvious solution without invoking Bayesian priors. Moreover, it is a tremendously useful idea in applications, so that Neyman's ideas on hypotheses testing and confidence intervals now play a major role in day-to-day applied statistics.

However, the success of the Neyman–Pearson lemma triggered new developments, leading to a more extreme form of statistical optimality that Fisher deeply distrusted. Even though Fisher's personal motives are suspect here, his philosophical qualms were far from groundless. Neyman's ideas, as later developed by Wald into decision theory, brought a qualitatively different spirit into statistics.

Fisher's maximum likelihood theory was launched in reaction to the rather shallow Laplacian Bayesianism of the previous century. Fisher's work demonstrated a more stringent approach to statistical inference. The Neyman–Wald decision theoretic school carried this spirit of astringency much further. A strict mathematical statement of the problem at hand, often phrased quite narrowly, followed by an optimal solution became the ideal. The practical result was a more sophisticated form of frequentist inference having enormous mathematical appeal.

Fisher, caught I think by surprise by this flanking attack from his right, complained that the Neyman–Wald decision theorists could be *accurate* without being *correct*. A favorite example of his concerned a Cauchy distribution with unknown center

$$f_\theta(x) = \frac{1}{\pi[1 + (x - \theta)^2]}. \tag{3}$$

Given a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from (3), decision theorists might try to provide the shortest interval of the form $\hat{\theta} \pm c$ that covers the true θ with probability 0.90. Fisher's objection, spelled out in his 1934 paper on ancillarity, was that c should be different for different samples $\mathbf{x}$ depending upon the correct amount of information in $\mathbf{x}$.

The decision theory movement eventually spawned its own counter-reformation. The neo-Bayesians, led by Savage and de Finetti, produced a more logical and persuasive Bayesianism, emphasizing subjective probabilities and personal decision making. In its most extreme form the Savage–de Finetti theory directly denies Fisher's claim of an impersonal logic of statistical inference. There has also been a postwar revival of interest in objectivist Bayesian theory, Laplacian in intent but based on Jeffreys's more sophisticated methods for choosing objective priors, which I shall talk more about later on.

Very briefly then, this is the way we arrived at the end of the 20th century with three competing philosophies of statistical inference: Bayesian; Neyman–Wald frequentist; and Fisherian. In many ways the Bayesian and frequentist philosophies stand at opposite poles from each other, with Fisher's ideas being somewhat of a compromise. I want to talk about that compromise next because it has a lot to do with the popularity of Fisher's methods.

4. THREE COMPETING PHILOSOPHIES

The chart in Figure 1 shows four major areas of disagreement between the Bayesians and the frequentists. These are not just philosophical disagreements. I chose the four categories because they lead to different behavior at the data-analytic level. For each category I have given a rough indication of Fisher's preferred position.

4.1 Individual Decision Making versus Scientific Inference

Bayes theory, and in particular Savage–de Finetti Bayesianism (the kind I'm focusing on here, though later I'll also talk about the Jeffreys brand of objec-

BAYES	FISHER	FREQUENTIST
1. Individual (personal decisions)	***	Universal (world of science)
2. Coherent (correct)	**********	Optimal (accurate)
3. Synthetic (combination)	****	Analytic (separation)
4. Optimistic (aggressive)	*****	Pessimistic (defensive)

FIG. 1. *Four major areas of disagreement between Bayesian and frequentist methods. For each one I have inserted a row of stars to indicate, very roughly, the preferred location of Fisherian inference.*

tive Bayesianism), emphasizes the individual decision maker, and it has been most successful in fields like business where individual decisions are paramount. Frequentists aim for universal acceptance of their inferences. Fisher felt that the proper realm of statistics was scientific inference, where it is necessary to persuade all or at least most of the world of science that you have reached the correct conclusion. Here Fisher is far over to the frequentist side of the chart (which is philosophically accurate but anachronistic, since Fisher's position predates both the Savage–de Finetti and Neyman–Wald schools).

4.2 Coherence versus Optimality

Bayesian theory emphasizes the coherence of its judgments, in various technical ways but also in the wider sense of enforcing consistency relationships between different aspects of a decision-making situation. Optimality in the frequentist sense is frequently incoherent. For example, the uniform minimum variance unbiased (UMVU) estimate of $\exp\{\theta\}$ does not have to equal $\exp\{\text{the UMVU of } \theta\}$, and more seriously there is no simple calculus relating the two different estimates. Fisher wanted to have things both ways, coherent and optimal, and in fact maximum likelihood estimation does satisfy

$$\exp\{\hat{\theta}\} = \widehat{\exp\{\theta\}}.$$

The tension between coherence and optimality is like the correctness–accuracy disagreement concerning the Cauchy example (3), where Fisher argued strongly for correctness. The emphasis on correctness, and a belief in the existence of a logic of statistical inference, moves Fisherian philosophy toward the Bayesian side of Figure 1. Fisherian practice is a less clear story. Different parts of the Fisherian program don't cohere with each other and in practice Fisher seemed quite willing to sacrifice logical consistency for a neat solution to a particular problem, for example, switching back and forth between frequentist and nonfrequentist justifications of the Fisher information. This kind of case-to-case expediency, which is a common attribute of modern data analysis has a frequentist flavor. I have located the Fisherian stars for this category a little closer to the Bayesian side of Figure 1, but spreading over a wide range.

4.3 Synthesis versus Analysis

Bayesian decision making emphasizes the collection of information across all possible sources, and the synthesis of that information into the final inference. Frequentists tend to break problems into separate small pieces that can be analyzed separately (and optimally). Fisher emphasized the use of all available information as a hallmark of correct inference, and in this way he is more in sympathy with the Bayesian position.

In this case Fisher tended toward the Bayesian position both in theory and in methodology: maximum likelihood estimation and its attendant theory of approximate confidence intervals based on Fisher information are superbly suited to the combination of information from different sources. (On the other hand, we have this quote from Yates and Mather: "In his own work Fisher was at his best when confronted with small self-contained sets of data. ... He was never much interested in the assembly and analysis of large amounts of data from varied sources bearing on a given issue." They blame this for his stubbornness on the smoking–cancer controversy. Here as elsewhere we will have to view Fisher as a lapsed Fisherian.)

4.4 Optimism versus Pessimism

This last category is more psychological than philosophical, but it is psychology rooted in the basic nature of the two competing philosophies. Bayesians tend to be more aggressive and risk-taking in their data analyses. There couldn't be a more pessimistic and defensive theory than minimax, to choose an extreme example of frequentist philosophy. It says that if anything can go wrong it will. Of course a minimax person might characterize the Bayesian position as "If anything can go right it will."

Fisher took a middle ground here. He scorns the finer mathematical concerns of the decision theorists ("Not only does it take a cannon to shoot a sparrow, but it misses the sparrow!"), but he fears averaging over the states of nature in a Bayesian way. One of the really appealing features of Fisher's work is its spirit of reasonable compromise, cautious but not overly concerned with pathological situations. This has always struck me as the right attitude toward most real-life problems, and it's certainly a large part of Fisher's dominance in statistical applications.

Looking at Figure 1, I think it is a mistake trying too hard to make a coherent philosophy out of Fisher's theories. From our current point of view they are easier to understand as a collection of extremely shrewd compromises between Bayesian and frequentist ideas. Fisher usually wrote as if he had a complete logic of statistical inference in hand, but that didn't stop him from changing his system when he thought up another landmark idea.

De Finetti, as quoted by Cifarelli and Regazzini, puts it this way: "Fisher's rich and manifold personality shows a few contradictions. His common

sense in applications on one hand and his lofty conception of scientific research on the other lead him to disdain the narrowness of a genuinely objectivist formulation, which he regarded as a *wooden attitude*. He professes his adherence to the objectivist point of view by rejecting the errors of the Bayes–Laplace formulation. What is not so good here is his mathematics, which he handles with mastery in individual problems but rather cavalierly in conceptual matters, thus exposing himself to clear and sometimes heavy criticism. From our point of view it appears probable that many of Fisher's observations and ideas are valid provided we go back to the intuitions from which they spring and free them from the arguments by which he thought to justify them."

Figure 1 describes Fisherian statistics as a compromise between the Bayesian and frequentist schools, but in one crucial way it is not a compromise: in its ease of use. Fisher's philosophy was always expressed in very practical terms. He seemed to think naturally in terms of computational algorithms, as with maximum likelihood estimation, analysis of variance and permutation tests. If anything is going to replace Fisher in the 21st century it will have to be a methodology that is equally easy to apply in day-to-day practice.

5. FISHER'S INFLUENCE ON CURRENT RESEARCH

There are three parts to this talk: past, present and future. The past part, which you have just seen, didn't do justice to Fisher's ideas, but the subject here is more one of influence than ideas, admitting of course that the influence is founded on the ideas's strengths. So now I am going to discuss Fisher's influence on current research.

What follows are several (actually six) examples of current research topics that have attracted a lot of attention recently. No claim of completeness is being made here. The main point I'm trying to make with these examples is that Fisher's ideas are still exerting a powerful influence on developments in statistical theory, and that this is an important indication of their future relevance. The examples will gradually get more speculative and futuristic, and will include some areas of development *not* satisfactorily handled by Fisher–holes in the Fisherian fabric—where we might expect future work to be more frequentist or Bayesian in motivation.

The examples will also allow me to talk about the new breed of applied problems statisticians are starting to see, the bigger, messier, more complicated data sets that we will have to deal with in the coming decades. Fisherian methods were fashioned to deal with the problems of the 1920s and 1930s. It is not a certainty that they will be equally applicable to the problems of the 21st century—a question I hope to shed at least a little light upon.

5.1 Fisher Information and the Bootstrap

This first example is intended to show how Fisher's ideas can pop up in current work, but be difficult to recognize because of computational advances. First, here is a very brief review of Fisher information. Suppose we observe a random sample $x_1, x_2, \ldots, x_n$ from a density function $f_\theta(x)$ depending on a single unknown parameter θ,

$$f_\theta(x) \to x_1, x_2, \ldots, x_n.$$

The Fisher information in any one x is the expected value of minus the second derivative of the log density,

$$i_\theta = \mathbf{E}_\theta\left\{-\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right\},$$

and the total Fisher information in the whole sample is ni_θ.

Fisher showed that the asymptotic standard error of the MLE is inversely proportional to the square root of the total information,

$$\mathrm{se}_\theta(\hat\theta) \doteq \frac{1}{\sqrt{ni_\theta}}, \tag{4}$$

and that no other consistent and sufficiently regular estimation of θ—essentially no other asymptotically, unbiased estimator—can do better.

A tremendous amount of philosophical interpretation has been attached to i_θ, concerning the meaning of statistical information, but in practice Fisher's formula (4) is most often used simply as a handy estimate of the standard error of the MLE. Of course, (4) by itself cannot be used directly because i_θ involves the unknown parameter θ. Fisher's tactic, which seems obvious but in fact is quite central to Fisherian methodology, is to *plug in* the MLE $\hat\theta$ for θ in (4), giving a usable estimate of standard error,

$$\widehat{\mathrm{se}} = \frac{1}{\sqrt{ni_{\hat\theta}}}. \tag{5}$$

Here is an example of formula (5) in action. Figure 2 shows the results of a small study designed to test the efficacy of an experimental antiviral drug.

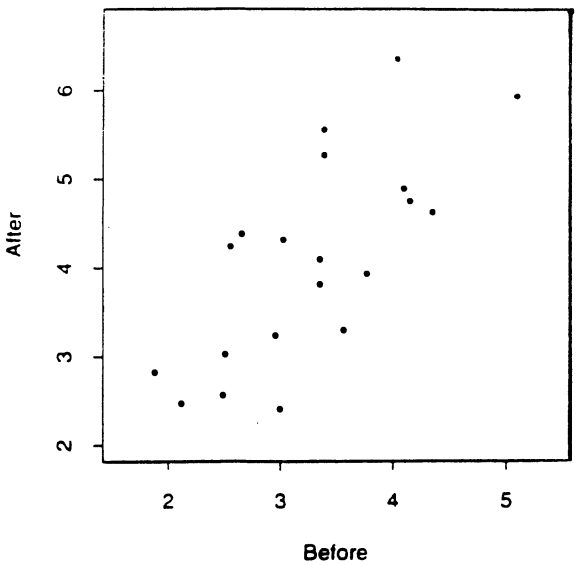

FIG. 2. *The cd4 data; 20 AIDS patients had their cd4 counts measured before and after taking an experimental drug; correlation coefficient* $\hat{\theta} = 0.723$.

A total of $n = 20$ AIDS patients had their cd4 counts measured before and after taking the drug, yielding data

$$x_i = (\text{before}_i, \text{after}_i) \quad \text{for } i = 1, 2, \ldots, 20.$$

The Pearson sample correlation coefficient was $\hat{\theta} = 0.723$. How accurate is this estimate?

If we assume a bivariate normal model for the data,

$$N_2(\mu, \Sigma) \to x_1, x_2, x_3, \ldots, x_{20}, \tag{6}$$

the notation indicating a random sample of 20 pairs from a bivariate normal distribution with expectation vector μ and covariance matrix Σ, then $\hat{\theta}$ is the MLE for the true correlation coefficient θ. The Fisher information for estimating θ turns out to be $i_\theta = 1/(1 - \theta^2)^2$ (after taking proper account of the "nuisance parameters" in (6)—one of those technical points I am avoiding in this talk) so (5) gives estimated standard error

$$\widehat{\text{se}} = \frac{(1 - \hat{\theta}^2)}{\sqrt{20}} = 0.107.$$

Here is a bootstrap estimate of standard error for the same problem, also assuming that the bivariate normal model is correct. In this context the bootstrap samples are generated from model (6), but with estimates $\hat{\mu}$ and $\hat{\Sigma}$ substituted for the unknown parameters μ and Σ:

$$N(\hat{\mu}, \hat{\Sigma}) \to x_1^*, x_2^*, x_3^*, \ldots, x_{20}^* \to \hat{\theta}^*,$$

where $\hat{\theta}^*$ is the sample correlation coefficient for the bootstrap data set $x_1^*, x_2^*, x_3^*, \ldots, x_{20}^*$.

This whole process was independently repeated 2,000 times, giving 2,000 bootstrap correlation coefficients $\hat{\theta}^*$. Figure 3 shows their histogram.

The empirical standard deviation of the 2,000 $\hat{\theta}^*$ values is

$$\widehat{\text{se}}_{\text{boot}} = 0.112,$$

which is the normal-theory bootstrap estimate of standard error for $\hat{\theta}$; 2,000 is 10 times more than needed for a standard error, but we will need all 2,000 later for the discussion of approximate confidence intervals.

5.2 The Plug-in Principle

The Fisher information and bootstrap standard error estimates, 0.107 and 0.112, are quite close to each other. This is no accident. Despite the fact that they look completely different, the two methods are doing very similar calculations. Both are using the "plug-in principle" as a crucial step in getting the answer.

Here is a plug-in description of the two methods:

- Fisher information—(i) compute an (approximate) formula for the standard error of the sample correlation coefficient as a function of the unknown parameters (μ, Σ); (ii) plug in estimates $(\hat{\mu}, \hat{\Sigma})$ for the unknown parameters (μ, Σ) in the formula;
- bootstrap—(i) plug in $(\hat{\mu}, \hat{\Sigma})$ for the unknown parameters (μ, Σ) in the mechanism generating the data; (ii) compute the standard error of the sample correlation coefficient, for the plugged-in mechanism, by Monte Carlo simulation.

The two methods proceed in reverse order, "compute and then plug in" versus "plug in and then compute," but this is a relatively minor technical difference. The crucial step in both methods, and the only statistical inference going on, is the substitution of the estimates $(\hat{\mu}, \hat{\Sigma})$ for the unknown parameters (μ, Σ), in other words the plug-in principle. Fisherian inference makes frequent use of the plug-in principle, and this is one of the main reasons that Fisher's methods are so convenient to use in practice. All possible inferential questions are answered by simply plugging in estimates, usually maximum likelihood estimates, for unknown parameters.

The Fisher information method involves cleverer mathematics than the bootstrap, but it has to because we enjoy a 10^7 computational advantage over Fisher. A year's combined computational effort by

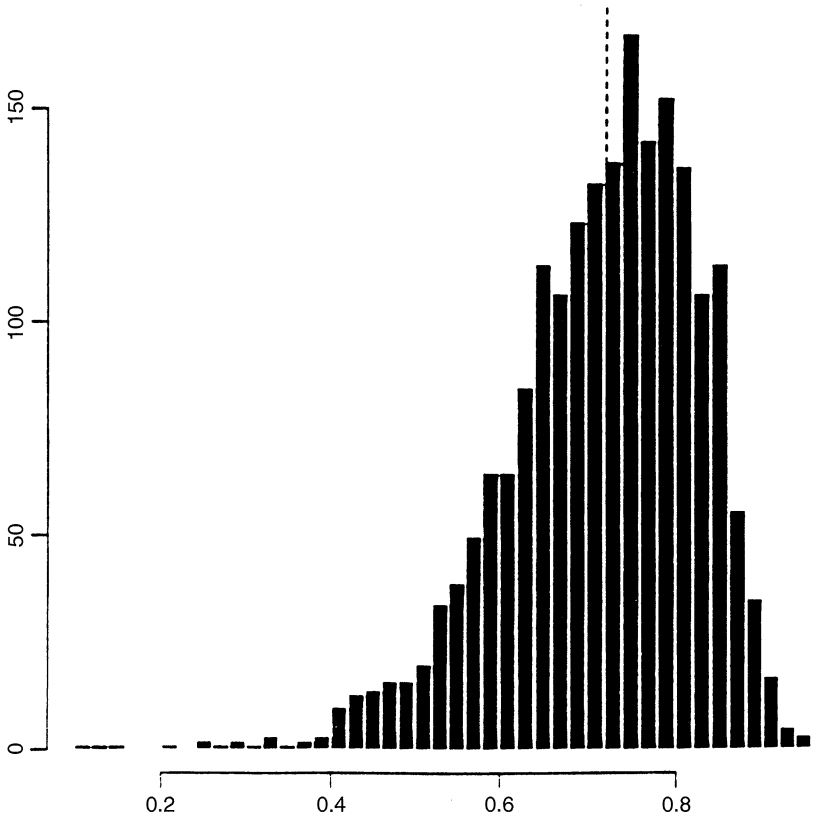

FIG. 3. *Histogram of* 2,000 *bootstrap correlation coefficients*; *bivariate normal sampling model.*

all the statisticians of 1925 wouldn't equal a minute of modern computer time. The bootstrap exploits this advantage to numerically extend Fisher's calculations to situations where the mathematics becomes hopelessly complicated. One of the less attractive aspects of Fisherian statistics is its overreliance on a small catalog of simple parametric models like the normal, understandable enough given the limitations of the mechanical calculators Fisher had to work with.

Modern computation has given us the opportunity to extend Fisher's methods to a much wider class of models, including nonparametric ones (the more usual arena of the bootstrap). We are beginning to see many such extensions, for example, the extension of discriminant analysis to CART, and the extension of linear regression to generalized additive models.

6. THE STANDARD INTERVALS

I want to continue the cd4 example, but proceeding from standard errors to confidence intervals. The confidence interval story illustrates how computer-based inference can be used to extend Fisher's ideas in a more ambitious way.

The MLE and its estimated standard error were used by Fisher to form approximate confidence intervals, which I like to call the *standard intervals* because of their ubiquity in day-to-day practice,

$$\hat{\theta} \pm 1.645\,\widehat{\mathrm{se}}. \tag{7}$$

The constant, 1.645, gives intervals of approximate 90% coverage for the unknown parameter θ, with 5% noncoverage probabilities at each end of the interval. We could use 1.96 instead of 1.645 for 95% coverage, and so on, but here I'll stick to 90%.

The standard intervals follow from Fisher's result that $\hat{\theta}$ is asymptotically normal, unbiased and with standard error fixed by the sample size and the Fisher information,

$$\hat{\theta} \to N(\theta, \mathrm{se}^2), \tag{8}$$

as in (4). We recognize (8) as one of Fisher's ideal "obvious" forms.

If usage determines importance then the standard intervals were Fisher's most important invention. Their popularity is due to a combination of optimality, or at least asymptotic optimality, with

computation tractability. The standard intervals are:

- *accurate*—their noncoverage probabilities, which are supposed to be 0.05 at each end of the interval, are actually

$$0.05 + c/\sqrt{n}\,, \tag{9}$$

  where c depends on the situation, so as the sample size n gets large we approach the nominal value 0.05 at rate $n^{-1/2}$;
- *correct*—the estimated standard error based on the Fisher information is the minimum possible for any asymptotically unbiased estimate of θ so interval (7) doesn't waste any information nor is it misleadingly optimistic;
- *automatic*—$\hat{\theta}$ and $\widehat{\text{se}}$ are computed from the same basic algorithm no matter how complicated the problem may be.

Despite these advantages, applied statisticians know that the standard intervals can be quite inaccurate in small samples. This is illustrated in the left panel of Figure 4 for the cd4 correlation example, where we see that the standard interval endpoints lie far to the right of the endpoints for the normal-theory exact 90% central confidence interval. In fact, we can see from the bootstrap histogram (reproduced from Figure 3) that in this case the asymptotic normality of the MLE hasn't taken hold at $n = 20$, so that there is every reason to doubt the standard interval. Being able to look at the histogram, which has a lot of information in it, is a luxury Fisher did not have.

Fisher suggested a fix for this specific situation: transform the correlation coefficient to $\hat{\phi} = \tanh^{-1}(\hat{\theta})$, that is, to

$$\hat{\phi} = \frac{1}{2}\log\frac{1+\hat{\theta}}{1-\hat{\theta}}, \tag{10}$$

apply the standard method on this scale and then transform the standard interval back to the θ scale. This was another one of Fisher's ingenious reduction methods. The $\tanh^{-1}$ transformation greatly accelerates convergence to normality, as we can see from the histogram of the 2,000 values of $\hat{\theta}^* = \tanh^{-1}(\hat{\theta}^*)$ in the right panel of Figure 4, and makes the standard intervals far more accurate. However, we have now lost the "automatic" property of the standard intervals. The $\tanh^{-1}$ transformation works only for the normal correlation coefficient and not for most other problems.

The standard intervals take literally the large sample approximation $\hat{\theta} \sim N(\theta, \text{se}^2)$, which says that $\hat{\theta}$ is normally distributed, is unbiased for θ and has a constant standard error. A more careful look at the asymptotics shows that each of these three assumptions can fail in a substantial way: the sampling distribution of $\hat{\theta}$ can be skewed; $\hat{\theta}$ can be biased as an estimate of θ; and its standard error can change with θ. Modern computation makes it practical to correct all three errors. I am going to mention two methods of doing so, the first using the bootstrap histogram, the second based on likelihood methods.

It turns out that there is enough information in the bootstrap histogram to correct all three errors

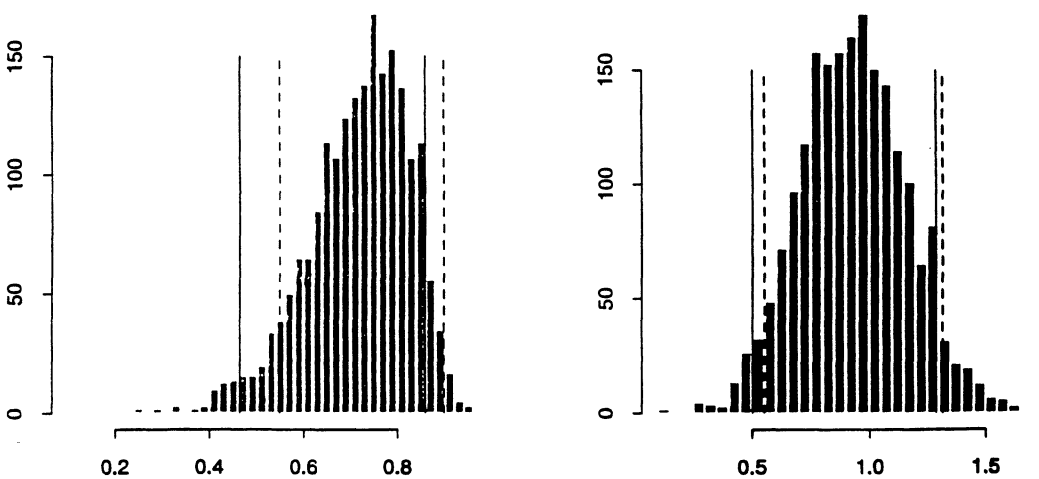

FIG. 4. *(Left panel) Endpoints of exact 90% confidence interval for cd4 correlation coefficient (solid lines) are much different than standard interval endpoints (dashed lines), as suggested by the nonnormality of the bootstrap histogram. (Right panel) Fisher's transformation normalizes the bootstrap histogram and makes the standard interval more accurate.*

of the standard intervals. The result is a system of approximate confidence intervals an order of magnitude more accurate, with noncoverage probabilities

$$0.05 + c/n$$

compared to (9), achieving what is called *second order accuracy*. Table 1 demonstrates the practical advantages of second order accuracy. In most situations we would not have exact endpoints as a "gold standard" for comparison, but second order accuracy would still point to the superiority of the bootstrap intervals.

The bootstrap method, and also the likelihood-based methods of the next section, are *transformation invariant*; that is, they give the same interval for the correlation coefficient whether or not you go through the $\tanh^{-1}$ transformation. In this sense they automate Fisher's wonderful transformation trick.

I like this example because it shows how a basic Fisherian construction, the standard intervals, can be extended by modern computation. The extension lets us deal easily with very complicated probability models, even nonparametric ones, and also with complicated statistics such as a coefficient in a stepwise robust regression.

Moreover, the extension is not just to a wider set of applications. Some progress in understanding the theoretical basis of approximate confidence intervals is made along the way. Other topics are springing up in the same fashion. For example, Fisher's 1925 work on the information loss for insufficient estimators has transmuted into our modern theories of the EM algorithm and Gibbs sampling.

7. CONDITIONAL INFERENCE, ANCILLARITY AND THE MAGIC FORMULA

Table 2 shows the occurrence of a very undesirable side effect in a randomized experiment that will be described more fully later. The treatment produces a smaller ratio of these undesirable effects than does the control, the sample log odds ratio being

$$\hat{\theta} = \log\left(\frac{1}{15}\Big/\frac{13}{3}\right) = -4.2.$$

TABLE 1
Endpoints of exact and approximate 90% confidence intervals for the cd4 correlation coefficient assuming bivariate normality

	Exact	Bootstrap	Standard
0.05	0.464	0.468	0.547
0.95	0.859	0.856	0.899

TABLE 2
The occurrence of adverse events in a randomized experiment; sample log odds ratio $\hat{\theta} = -4.2$

	Yes	No	
Treatment	1	15	16
Control	13	3	16
	14	18	

Fisher wondered how one might make appropriate inferences for θ, the true log odds ratio. The trouble here is nuisance parameters. A multinomial model for the 2×2 table has three free parameters, representing four cell probabilities constrained to add up to 1, and in some sense two of the three parameters have to be eliminated in order to get at θ. To do this Fisher came up with another device for reducing a complicated situation to a simple form.

Fisher showed that if we condition on the marginals of the table, then the conditional density of $\hat{\theta}$ given the marginals depends only θ. The nuisance parameters disappear. This conditioning is "correct" he argued because the marginals are acting as what might be called *approximate ancillary statistics*. That is, they do not carry much direct information concerning the value of θ, but they have something to say about how accurately $\hat{\theta}$ estimates θ. Later Neyman gave a much more specific frequentist justification for conditioning on the marginals, through what is now called *Neyman structure*.

For the data in Table 2, the conditional distribution of $\hat{\theta}$ given the marginals yields $[-6.3, -2.4]$ as a 90% confidence interval for θ, ruling out the null hypothesis value $\theta = 0$ where Treatment equals Control. However, the conditional distribution is not easy to calculate, even in this simple case, and it becomes prohibitive in more complicated situations.

In his 1934 paper, which was the capstone of Fisher's work on efficient estimation, he solved the conditioning problem for translation families. Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a random sample from a Cauchy distribution (3) and that we wish to use $\mathbf{x}$ to make inferences about θ, the unknown center point of the distribution. In this case there is a genuine ancillary statistic $\mathbf{A}$, the vector of spacings between the ordered values of $\mathbf{x}$. Again Fisher argued that correct inferences about θ should be based on $f_\theta(\hat{\theta}|\mathbf{A})$, the conditional density of the MLE $\hat{\theta}$ given the ancillary $\mathbf{A}$, not on the unconditional density $f_\theta(\hat{\theta})$.

Fisher also provided a wonderful trick for calculating $f_\theta(\hat{\theta}|\mathbf{A})$. Let $L(\theta)$ be the likelihood function:

the unconditional density of the whole sample, considered as a function of θ with $\mathbf{x}$ fixed. Then it turns out that

$$(11) \qquad f_\theta(\hat{\theta}|\mathbf{A}) = c\frac{L(\theta)}{L(\hat{\theta})},$$

where c is a constant. Formula (11) allows us to compute the conditional density $f_\theta(\hat{\theta}|\mathbf{A})$ from the likelihood, which is easy to calculate. It also hints at a deep connection between likelihood-based inference, a Fisherian trademark, and frequentist methods.

Despite this promising start, the promise went unfulfilled in the years following 1934. The trouble was that formula (11) applies only in very special circumstances, not including the 2×2 table example, for instance. Recently, though, there has been a revival of interest in likelihood-based conditional inference. Durbin, Barndorff-Nielsen, Hinkley and others have developed a wonderful generalization of (11) that applies to a wide variety of problems having approximate ancillaries, the so-called magic formula

$$(12) \qquad f_\theta(\hat{\theta}|\mathbf{A}) = c\frac{L(\theta)}{L(\hat{\theta})}\left\{-\frac{d^2}{d\theta^2}\log L(\theta)|_{\theta=\hat{\theta}}\right\}^{1/2}.$$

The bracketed factor is constant in the Cauchy situation, reducing (12) back to (11).

Likelihood-based conditional inference has been pushed forward in current work by Fraser, Cox and Reid, McCullagh, Barndorff-Nielson, Pierce, DiCiccio and many others. It represents a major effort to perfect and extend Fisher's goal of an inferential system based directly on likelihoods.

In particular the magic formula can be used to generate approximate confidence intervals that are more accurate than the standard intervals, at least second order accurate. These intervals agree to second order with the bootstrap intervals. If this were not true, then one or both of them would not be second order correct. Right now it looks like attempts to improve upon the standard intervals are converging from two directions: likelihood and bootstrap.

Results like (12) have enormous potential. Likelihood inference is the great unfulfilled promise of Fisherian statistics—the promise of a theory that directly interprets likelihood functions in a way that simultaneously satisfies Bayesians and frequentists. Fulfilling that promise, even partially, would greatly influence the shape of 21st century statistics.

8. FISHER'S BIGGEST BLUNDER

Now I'll start edging gingerly into the 21st century by discussing some topics where Fisher's ideas have not been dominant, but where they might or might not be important in future developments. I am going to begin with the fiducial distribution, generally considered to be Fisher's biggest blunder. But in Arthur Koestler's words "The history of ideas is filled with barren truths and fertile errors." If fiducial inference is an error it certainly has been a fertile one.

In terms of Figure 1, the Bayesian–frequentist comparison chart, fiducial inference was Fisher's closest approach to the Bayesian side of the ledger. Fisher was trying to codify an objective Bayesianism in the Laplace tradition but without using Laplace's ad hoc uniform prior distributions. I believe that Fisher's continuing devotion to fiducial inference had two major influences, a negative reaction against Neyman's ideas and a positive attraction to Jeffreys's point of view.

The solid line in Figure 5 is the fiducial density for a binomial parameter θ having observed 3 successes in 10 trials,

$$s \sim \text{Binomial}(n, \theta), \quad s = 3 \text{ and } n = 10.$$

Also shown is an approximate fiducial density that I will refer to later. Fisher's fiducial theory at its boldest treated the solid curve as a genuine a posteriori density for θ even though, or perhaps because, no prior assumptions had been made.

8.1 The Confidence Density

We could also call the fiducial distribution the "confidence density" because this is an easy way to motivate the fiducial construction. As I said earlier, Fisher would have hated this name.

Suppose that for every value of α between 0 and 1 we have an upper 100 αth confidence limit $\hat{\theta}[\alpha]$ for θ, so that by definition

$$\text{prob}\{\theta < \hat{\theta}[\alpha]\} = \alpha.$$

We can interpret this as a probability distribution for θ given the data if we are willing to accept the classic *wrong* interpretation of confidence,

> θ is in the interval $(\hat{\theta}[0.90], \hat{\theta}[0.91])$
> with probability 0.01, and so on.

Going to the continuous limit gives the "confidence density," a name Neyman would have hated.

The confidence density *is* the fiducial distribution, at least in those cases where Fisher would have considered the confidence limits to be inferen-

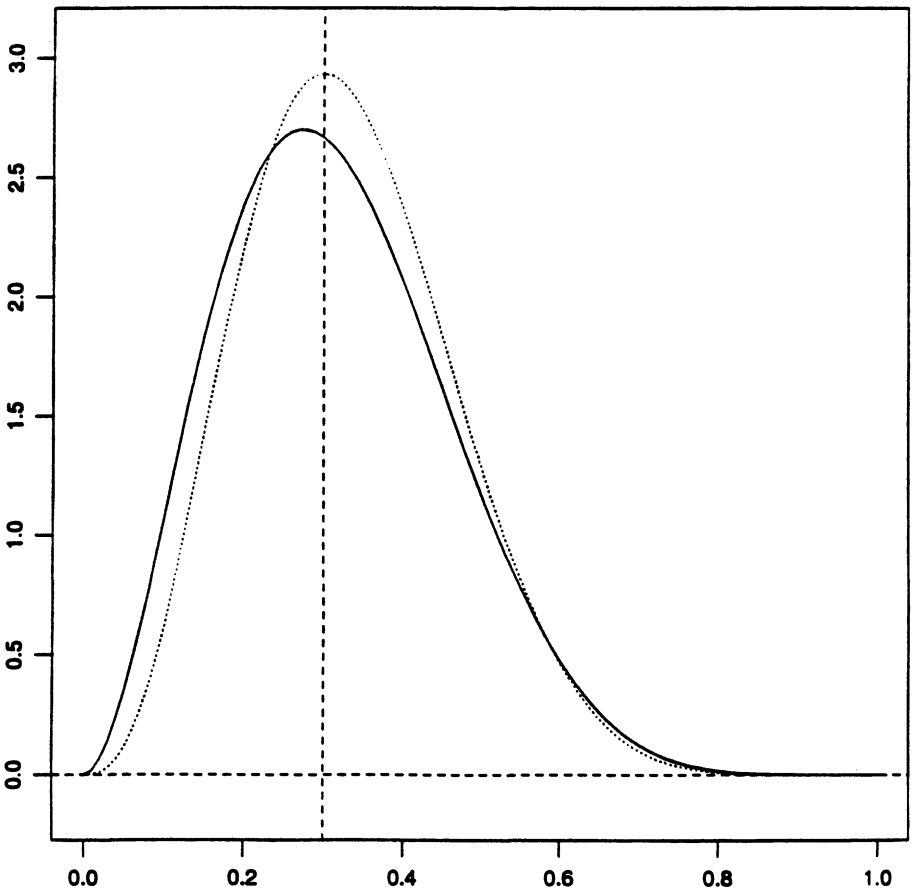

FIG. 5. *Fiducial density for a binomial parameter θ having observed 3 successes out of 10 trials. The dashed line is an approximation that is useful in complicated situations.*

tially correct. The fiducial distribution in Figure 5 is the confidence density based on the usual confidence limits for θ (taking into account the discrete nature of the binomial distribution): $\hat{\theta}[\alpha]$ is the value of θ such that $S \sim \text{Binomial}(10, \theta)$ satisfies

$$\text{prob}\{S > 3\} + \tfrac{1}{2}\,\text{prob}\{S = 3\} = \alpha.$$

Fisher was uncomfortable applying fiducial arguments to discrete distributions because of the ad hoc continuity corrections required, but the difficulties caused are more theoretical than practical.

The advantage of stating fiducial ideas in terms of the confidence density is that they then can be applied to a wider class of problems. We can use the approximate confidence intervals mentioned earlier, either the bootstrap or the likelihood ones, to get approximate fiducial distribution even in very complicated situations having lots of nuisance parameters. (The dashed curve in Figure 5 is the confidence density based on approximate bootstrap intervals.) And there are practical reasons why it would be very convenient to have good approximate fiducial distributions, reasons connected with our profession's 250-year search for a dependable objective Bayes theory.

8.2 Objective Bayes

By "objective Bayes" I mean a Bayesian theory in which the subjective element is removed from the choice of prior distribution; in practical terms a universal recipe for applying Bayes theorem in the absence of prior information. A widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance.

I have in mind here dealing with messy, complicated problems where we are trying to combine information from disparate sources—doing a metaanalysis, for example. Bayesian methods are particularly well-suited to such problems. This is particularly true now that techniques like the Gibbs sampler and Markov chain Monte Carlo are available for integrating the nuisance parameters out of high-dimensional posterior distributions.

The trouble of course is that the statistician still has to choose a prior distribution in order to use

Bayes's theorem. An unthinking use of uniform priors is no better now than it was in Laplace's day. A lot of recent effort has been put into the development of uninformative or objective prior distributions, priors that eliminate nuisance parameters safely while remaining neutral with respect to the parameter of interest. Kass and Wasserman's 1996 *JASA* article reviews current developments by Berger, Bernardo and many others, but the task of finding genuinely objective priors for high-dimensional problems remains daunting.

Fiducial distributions, or confidence densities, offer a way to finesse this difficulty. A good argument can be made that the confidence density is the posterior density for the parameter of interest, after all of the nuisance parameters have been integrated out in an objective way. If this argument turns out to be valid, then our progress in constructing approximate confidence intervals, and approximate confidence densities, could lead to an easier use of Bayesian thinking in practical problems.

This is all quite speculative, but here is a safe prediction for the 21st century: statisticians will be asked to solve bigger and more complicated problems. I believe that there is a good chance that objective Bayes methods will be developed for such problems, and that something like fiducial inference will play an important role in this development. Maybe Fisher's biggest blunder will become a big hit in the 21st century!

9. MODEL SELECTION

Model selection is another area of statistical research where important developments seem to be building up, but without a definitive breakthrough. The question asked here is how to select the model itself, not just the continuous parameters of a given model, from the observed data. F-tests, and "F" stands for Fisher, help with this task, and are certainly the most widely used model selection techniques. However, even in relatively simple problems things can get complicated fast, as anyone who has gotten lost in a tangle of forward and backward stepwise regression programs can testify.

The fact is that classic Fisherian estimation and testing theory are a good start, but not much more than that, on model selection. In particular, maximum likelihood estimation theory and model fitting do not account for the number of free parameters being fit, and that is why frequentist methods like Mallow's C_p, the Akaike information criterion and cross-validation have evolved. Model selection seems to be moving away from its Fisherian roots.

Now statisticians are starting to see really complicated model selection problems, with thousands and even millions of data points and hundreds of candidate models. A thriving area called machine learning has developed to handle such problems, in ways that are not yet very well connected to statistical theory.

Table 3, taken from Gail Gong's 1982 thesis, shows part of the data from a model selection problem that is only moderately complicated by today's standards, though hopelessly difficult from a prewar viewpoint. A "training set" of 155 chronic hepatitis patients were measured on 19 diagnostic prediction variables. The outcome variable y was whether or not the patient died from liver failure (122 lived, 33 died), the goal of the study being to develop a prediction rule for y in terms of the diagnostic variables.

In order to predict the outcome, a logistic regression model was built up in three steps:

- Individual logistic regressions were run for each of the 19 predictors, yielding 13 that were significant at the 0.05 level.
- A forward stepwise logistic regression program, including only those patients with none of the 13 predictors missing, retained 5 of the 13 predictors at significance level 0.10.
- A second forward stepwise logistic regression program, including those patients with none of the 5 predictors missing, retained 4 of the 5 at significance level 0.05.

These last four variables,

(13) ascites, (15) bilirubin,
(7) malaise, (20) histology,

were deemed the "important predictors." The logistic regression based on them misclassified 16% of the 155 patients, with cross-validation suggesting a true error rate of about 20%.

A crucial question concerns the validity of the selected model. Should we take the four "important predictors" very seriously in a medical sense? The bootstrap answer seems to be "probably not," even though it was natural for the medical investigator to do so given the impressive amount of statistical machinery involved in their selection.

Gail Gong resampled the 155 patients, taking as a unit each patient's entire record of 19 predictors and response. For each bootstrap data set of 155 resampled records, she reran the three-stage logistic regression model, yielding a bootstrap set of "important predictors." This was done 500 times. Figure 6 shows the important predictors for the final 25 bootstrap data sets. The first of these is (13, 7, 20, 15), agreeing except for order with the set (13, 15, 7, 20) from the original data. This didn't happen in any other of the 499 bootstrap cases. In

TABLE 3

155 chronic hepatitis patients were measured on 19 diagnostic variables; data shown for the last 11 patients; outcome $y = 0$ or 1 as patient lived or died; negative numbers indicate missing data

y	Constant 1	Age 2	Sex 3	Steroid 4	Antiviral 5	Fatigue 6	Malaise 7	Anorexia 8	Liver Big 9	Liver Firm 10	Spleen Palp 11	Spiders 12	Ascites 13	Varices 14	Bilirubin 15	Alk Phos 16	SGOT 17	Albumin 18	Protein 19	Histology 20	#
1	1	45	1	2	2	1	1	1	2	2	2	1	1	2	1.90	−1	114	2.4	−1	−3	145
0	1	31	1	1	2	1	2	2	2	2	2	2	2	2	1.20	75	193	4.2	54	2	146
1	1	41	1	2	2	1	2	2	2	1	1	1	2	1	4.20	65	120	3.4	−1	−3	147
1	1	70	1	1	2	1	1	1	−3	−3	−3	−3	−3	−3	1.70	109	528	2.8	35	2	148
0	1	20	1	1	2	2	2	2	2	−3	2	2	2	2	0.90	89	152	4.0	−1	2	149
0	1	36	1	2	2	2	2	2	2	2	2	2	2	2	0.60	120	30	4.0	−1	2	150
1	1	46	1	2	2	1	1	1	2	2	2	1	1	1	7.60	−1	242	3.3	50	−3	151
0	1	44	1	2	2	1	2	2	2	1	2	2	2	2	0.90	126	142	4.3	−1	2	152
0	1	61	1	1	2	1	1	2	1	1	2	1	2	2	0.80	95	20	4.1	−1	2	153
0	1	53	2	1	2	1	2	2	2	2	1	1	2	1	1.50	84	19	4.1	48	−3	154
1	1	43	1	2	2	1	2	2	2	2	1	1	1	2	1.20	100	19	3.1	42	2	155

13	7	20	15			
13	19	6				
20	16	19				
20	19					
14	18	7	16	2		
18	20	7	11			
20	19	15				
20						
13	12	15	8	18	7	19
15	13	19				
13	4					
12	15	3				
15	16	3				
15	20	4				
16	13	2	19			
18	20	3				
13	15	20				
15	13					
15	20	7				
13						
15						
13	14					
12	20	18				
2	20	15	7	19	12	
13	20	15	19			

FIG. 6. *The set of "important predictors" selected in the last 25 of 500 bootstrap replications of the three-step logistic regression model selection program; original choices were* (13, 15, 7, 20).

all 500 bootstrap replications only variable 20, histology, which appeared 295 times, was "important" more than half of the time. These results certainly discourage confidence in the causal nature of the predictor variables (13, 15, 7, 20).

Or do they? It seems like we should be able to use the bootstrap results to quantitatively assess the validity of the various predictors. Perhaps they could also help in selecting a better prediction model. Questions like these are being asked these days, but the answers so far are more intriguing than conclusive.

It is not clear to me whether Fisherian methods will play much of a role in the further progress of model selection theory. Figure 6 makes model selection look like an exercise in discrete estimation, while Fisher's MLE theory was always aimed at continuous situations. Direct frequentist methods like cross-validation seem more promising right now, and there have been some recent developments in Bayesian model selection, but in fact our best efforts so far are inadequate for problems like the hepatitis data. We could badly use a clever Fisherian trick for reducing complicated model selection problems to simple obvious ones.

10. EMPIRICAL BAYES METHODS

As a final example, I wanted to say a few words about empirical Bayes methods. Empirical Bayes

seems like the wave of the future to me, but it seemed that way 25 years ago and the wave still hasn't washed in, despite the fact that it is an area of enormous potential importance. It is not a topic that has had much Fisherian input.

Table 4 shows the data for an empirical Bayes situation: independent clinical trials were run in 41 cities, comparing the occurrence of recurrent bleeding, an undesirable side effect, for two stomach ulcer surgical techniques, a new treatment and an older control. Each trial yielded an estimate of the true log odds ratio for recurrent bleeding, Treatment versus Control,

$$\theta_i = \text{log odds ratio in city } i, \quad i = 1, 2, \ldots, 41.$$

In city 8, for example, we have the estimate seen earlier in Table 2,

$$\hat{\theta} = \log\left(\frac{1}{15}\Big/\frac{13}{3}\right) = -4.2,$$

indicating that the new surgery was very effective in reducing recurrent bleeding, at least in city 8.

Figure 7 shows the likelihoods for θ_i in 10 of the 41 cities. These are conditional likelihoods, using Fisher's trick of conditioning on the marginals to get rid of the nuisance parameters in each city. It seems clear that the log odds ratios θ_i are not all the same. For instance, the likelihoods for cities 8 and 13 barely overlap. On the other hand, the θ_i values are not wildly discrepant, most of the 41 likelihood functions concentrating themselves on the range $(-6, 3)$. (This is the kind of complicated inferential situation I was worrying about in the discussion of fiducial inference, confidence densities and objective Bayes methods.)

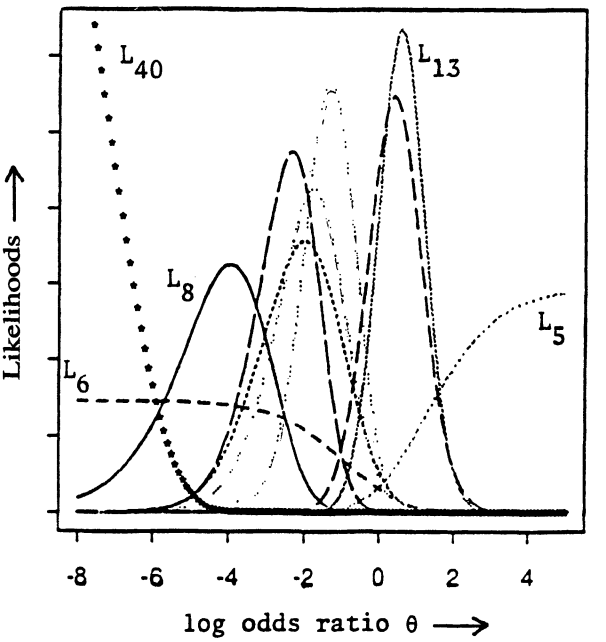

FIG. 7. *Individual likelihood functions for θ_i, for 10 of the 41 experiments in Table 4; L_8, the likelihood for the log odds ratio in city 8, lies to the left of most of the others.*

TABLE 4

Ulcer data: 41 independent experiments concerning the number of occurrences of recurrent bleeding following ulcer surgery; (a, b) = (# bleeding; # nonbleeding) for Treatment, a new surgical technique; (c, d) is the same for Control, an older surgery; $\hat{\theta}$ is the sample log odds ratio, with estimated standard deviation $\widehat{\text{SD}}$; stars indicate cases shown in Figure 7

Experiment	a	b	c	d	$\hat{\theta}$	$\widehat{\text{SD}}$	Experiment	a	b	c	d	$\hat{\theta}$	$\widehat{\text{SD}}$
1*	7	8	11	2	−1.84	0.86	21	6	34	13	8	−2.22	0.61
2	8	11	8	8	−0.32	0.66	22	4	14	5	34	66	0.71
3*	5	29	4	35	0.41	0.68	23	14	54	13	61	20	0.42
4	7	29	4	27	0.49	0.65	24	6	15	8	13	−43	0.64
5*	3	9	0	12	inf	1.57	25	0	6	6	0	−inf	2.08
6*	4	3	4	0	−inf	1.65	26	1	9	5	10	−1.50	1.02
7*	4	13	13	11	−1.35	0.68	27	5	12	5	10	−0.18	0.73
8*	1	15	13	3	−4.17	1.04	28	0	10	12	2	−inf	1.60
9	3	11	7	15	−0.54	0.76	29	0	22	8	16	−inf	1.49
10*	2	36	12	20	−2.38	0.75	30	2	16	10	11	−1.98	0.80
11	6	6	8	0	−inf	1.56	31	1	14	7	6	−2.79	1.01
12*	2	5	7	2	−2.17	1.06	32	8	16	15	12	−0.92	0.57
13*	9	12	7	17	0.60	0.61	33	6	6	7	2	−1.25	0.92
14	7	14	5	20	0.69	0.66	34	0	20	5	18	−inf	1.51
15	3	22	11	21	−1.35	0.68	35	4	13	2	14	0.77	0.87
16	4	7	6	4	−0.97	0.86	36	10	30	12	8	−1.50	0.57
17	2	8	8	2	−2.77	1.02	37	3	13	2	14	0.48	0.91
18	1	30	4	23	−1.65	0.98	38	4	30	5	14	−0.99	0.71
19	4	24	15	16	−1.73	0.62	39	7	31	15	22	−1.11	0.52
20	7	36	16	27	−1.11	0.51	40*	0	34	34	0	−inf	2.01
							41	0	9	0	16	NA	2.04

Notice that L_8, the likelihood for θ_8, lies to the left of most of the other curves. This would still be true if we could see all 41 curves instead of just 10 of them. In other words, θ_8 appears to be more negative than the log odds ratios in most of the other cities.

What is a good estimate or confidence interval for θ_8? Answering this question depends on how much the results in other cities influence our thinking about city 8. That is where empirical Bayes theory comes in, giving us a systematic framework for combining the direct information for θ_8 from city 8's experiment with the indirect information from the experiments in the other 40 cities.

The ordinary 90% confidence interval for θ_8, based only on the data $(1, 15, 13, 3)$ from its own experiment, is

$$\theta_8 \in [-6.3, -2.4]. \tag{13}$$

Empirical Bayes methods give a considerably different result. The empirical Bayes analysis uses the data in the other 40 cities to estimate a prior density for log odds ratios. This prior density can be combined with the likelihood L_8 for city 8, using Bayes theorem, to get a central 90% a posteriori interval for θ_8,

$$\theta_8 \in [-5.1, -1.8]. \tag{14}$$

The fact that most of the cities had less negatively tending results than city 8 plays an important role in the empirical Bayes analysis. The Bayesian prior estimated from the other 40 cities says that θ_8 is unlikely to be as negative as its own data by itself would indicate.

The empirical Bayes analysis implies that there is a lot of information in the other 40 cities's data for estimating θ_8, as a matter of fact, just about as much as in city 8's own data. This kind of "other" information does not have a clear Fisherian interpretation. The whole empirical Bayes analysis is heavily Bayesian, as if we had begun with a genuinely informative prior for θ_8 and yet it still has some claims to frequentist objectivity.

Perhaps we are verging here on a new compromise between Bayesian and frequentist methods, one that is fundamentally different from Fisher's proposals. If so, the 21st century could look a lot less Fisherian, at least for problems with parallel structure like the ulcer data. Right now there aren't many such problems. This could change quickly if the statistics community became more confident about analyzing empirical Bayes problems. There weren't many factorial design problems before Fisher provided an effective methodology for handling them. Scientists tend to bring us the problems we can solve. The current attention to metaanalysis and hierarchical models certainly suggests a growing interest in the empirical Bayes kind of situation.

11. THE STATISTICAL TRIANGLE

The development of modern statistical theory has been a three-sided tug of war between the Bayesian, frequentist and Fisherian viewpoints. What I have been trying to do with my examples is apportion the influence of the three philosophies on several topics of current interest: standard error estimation; approximate confidence intervals; conditional inference; objective Bayes theories and fiducial inference; model selection; and empirical Bayes techniques.

Figure 8, the statistical triangle, does this more concisely. It uses barycentric coordinates to indicate the influence of Bayesian, frequentist and Fisherian thinking upon a variety of active research areas. The Fisherian pole of the triangle is located between the Bayesian and frequentist poles, as in Figure 1, but here I have allocated Fisherian philosophy its own dimension to take account of its distinctive operational features: reduction to "obvious" types; the plug-in principle; an emphasis on inferential correctness; the direct interpretation of likelihoods; and the use of automatic computational algorithms.

Of course, a picture like this cannot be more than roughly accurate, even if one accepts the author's prejudices, but many of the locations are difficult to argue with. I had no trouble placing conditional inference and partial likelihood near the Fisherian pole, robustness at the frequentist pole and multiple imputation near the Bayesian pole. Empirical Bayes is clearly a mixture of Bayesian and frequentist ideas. Bootstrap methods combine the convenience of the plug-in principle with a strong frequentist desire for accurate operating characteristics, particularly for approximate confidence intervals, while the jackknife's development has been more purely frequentistic.

Some of the other locations in Figure 8 are more problematical. Fisher provided the original idea behind the EM algorithm, and in fact the self-consistency of maximum likelihood estimation (when missing data is filled in by the statistician) is a classic Fisherian correctness argument. On the other hand EM's modern development has had a strong Bayesian component, seen more clearly in the related topic of Gibbs sampling. Similarly, Fisher's method for combining independent p-values is an early form of metaanalysis, but the subject's recent growth has been strongly frequentist.

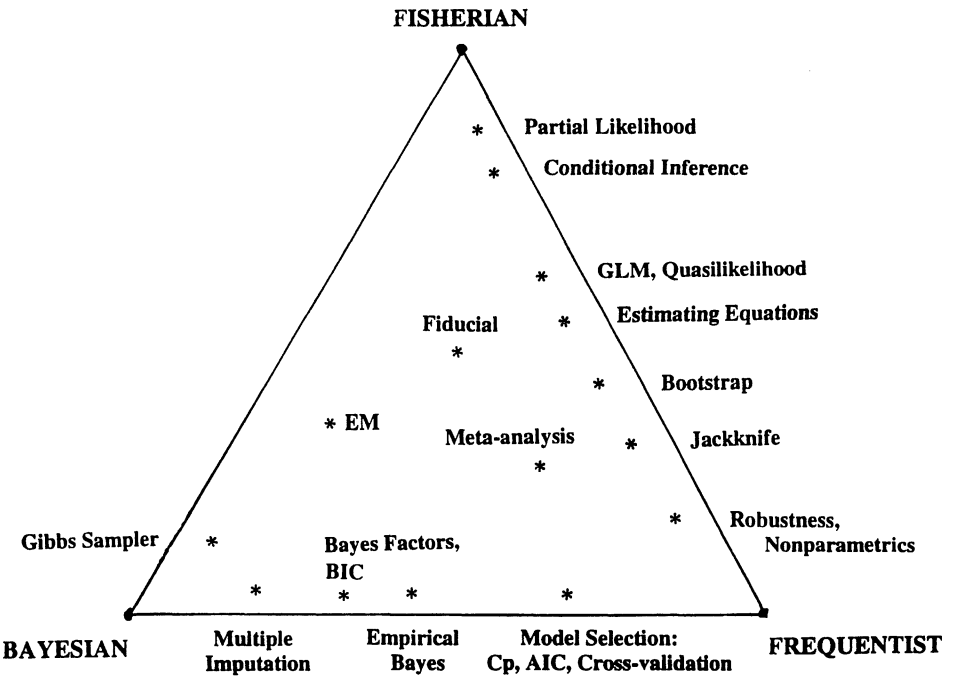

FIG. 8. *A barycentric picture of modern statistical research, showing the relative influence of the Bayesian, frequentist and Fisherian philosophies upon various topics of current interest.*

The trouble here is that Fisher wasn't always a Fisherian, so it is easy to confuse parentage with development.

The most difficult and embarrassing case concerns what I have been calling "objective Bayes" methods, among which I included fiducial inference. One definition of frequentism is the desire to do well, or at least not to do poorly, against every possible prior distribution. The Bayesian spirit, as epitomized by Savage and de Finetti, is to do very well against one prior distribution, presumably the right one.

There have been a variety of objective Bayes compromises between these two poles. Working near the frequentist end of the spectrum, Welch and Peers showed how to calculate priors whose a posteriori credibility intervals coincide closely with standard confidence intervals. Jeffreys's work, which has led to vigorous modern development of Bayesian model selection, is less frequentistic. In a bivariate normal situation Jeffreys would recommend the same prior distribution for estimating the correlation coefficient or for the ratio of expectations, while the Welch–Peers theory would use two different priors in order to separately match each of the frequentist solutions.

Nevertheless Jeffreys's Bayesianism has an undeniable objectivist flavor. Erich Lehmann (personal communication) had this to say: "If one separates the two Bayesian concepts [Savage–de Finetti and Jeffreys] and puts only the subjective version in your Bayesian corner, it seems to me that something interesting happens: the Jeffreys concept moves to the right and winds up much closer to the frequency corner than to the Bayesian one. For example, you contrasted Bayes as optimistic and risk-taking with frequentist as pessimistic and playing it safe. On both of these scales Jeffreys is much closer to the frequentist end of the spectrum. In fact, the concept of uninformative prior is philosophically close to Wald's least favorable distribution, and the two often coincide."

Lehmann's advice is followed a bit in Figure 8, where the Bayesian model selection (BIC) point, a direct legacy of Jeffreys's work, has been moved a little ways toward the frequentist pole. However, I have located fiducial inference, Fisher's form of objective Bayesianism, near the center of the triangle. There isn't much work in that area right now but there is a lot of demand coming from all three directions.

The point of my examples, and the main point of this talk, was to show that Fisherian statistics, is not a dead language and that it continues to inspire new research. I think this is clear in Figure 8, even allowing for its inaccuracies. But Fisher's language is not the only language in town, and it is not even the dominant language of our research journals. That prize would have to go to a rather casual frequentism, not usually as hard-edged as pure decision theory these days. We might ask what Figure 8 will look like 20 or 30 years from now, and

whether there will be many points of active research interest lying near the Fisherian pole of the triangle.

12. R. A. FISHER IN THE 21ST CENTURY

Most talks about the future are really about the present, and this one has certainly been no exception. But here at the end of the talk, and nearly at the end of the 20th century, we can peek cautiously ahead and speculate at least a little bit about the future of Fisherian statistics.

Of course Fisher's fundamental discoveries like sufficiency, Fisher information, the asymptotic efficiency of the MLE, experimental design and randomization inference are not going to disappear. They might become less visible though. Right now we use those ideas almost exactly as Fisher coined them, but modern computing equipment could change that.

For example, maximum likelihood estimates can be badly biased in certain situations involving a great many nuisance parameters (as in the Neyman–Scott paradox.) A computer-modified version of the MLE that was less less biased could become the default estimator of choice in applied problems. REML estimation of variance components offers a current example. Likewise, with the universal spread of high-powered computers statisticians might automatically use some form of the more accurate confidence intervals I mentioned earlier instead of the standard intervals.

Changes like these would conceal Fishers's influence, but not really diminish it. There are a couple of good reasons though that one might expect more dramatic changes in the statistical world, the first of these being the miraculous improvement in our computational equipment, by orders of magnitude every decade. Equipment is destiny in science, and statistics is no exception to that rule. Second, statisticians are being asked to solve bigger, harder, more complicated problems, under such names as pattern recognition, DNA screening, neural networks, imaging and machine learning. New problems have always evoked new solutions in statistics, but this time the solutions might have to be quite radical ones.

Almost by definition it's hard to predict radical change, but I thought I would finish with a few speculative possibilities about a statistical future that might, or might not, be a good deal less Fisherian.

12.1 A Bayesian World

In 1974 Dennis Lindley predicted that the 21st century would be Bayesian. (I notice that his recent *Statistical Science* interview now predicts the year 2020.) He could be right. Bayesian methods are attractive for complicated problems like the ones just mentioned, but unless the scientific world changes the way it thinks I can't imagine subjective Bayes methods taking over. What I called objective Bayes, the use of neutral or uninformative priors, seems a lot more promising and is certainly in the air these days.

A successful objective Bayes theory would have to provide good frequentist properties in familiar situations, for instance, reasonable coverage probabilities for whatever replaces confidence intervals. Such a Bayesian world might not seem much different than the current situation except for more straightforward analyses of complicated problems like multiple comparisons. One can imagine the statistician of the year 2020 hunched over his or her supercomputer terminal, trying to get Proc Prior to run successfully, and we can only wish that future colleague "good luck."

12.2 Nonparametrics

As part of our Fisherian legacy we tend to overuse simple parametric models like the normal. A nonparametric world, where parametric models were a last resort instead of the first, would favor the frequentist vertex of the triangle picture.

12.3 A New Synthesis

The postwar years and especially the last few decades have been more notable for methodological progress than the development of fundamental new ideas in the theory of statistical inference. This doesn't mean that such developments are finished forever. Fisher's work came out of the blue in the 1920s, and maybe our field is due for another bolt of lightening.

It's easy for us to imagine that Fisher, Neyman and the others were lucky to live in a time when all the good ideas hadn't been plucked from the trees. In fact, we are the ones living in the golden age of statistics—the time when computation has become fast and easy. In this sense we are overdue for a new statistical paradigm, to consolidate the methodological gains of the postwar period. The rubble is building up again, to use Joan Fisher Box's simile, and we could badly use a new Fisher to put our world in order.

My actual guess is that the old Fisher will have a very good 21st century. The world of applied statistics seems to need an effective compromise between Bayesian and frequentist ideas, and right now there is no substitute in sight for the Fisherian synthesis. Moreover, Fisher's theories are well suited to life in

the computer age. Fisher seemed naturally to think in algorithmic terms. Maximum likelihood estimates, the standard intervals, ANOVA tables, permutation tests are all expressed algorithmically and are easy to extend with modern computation.

Let me say finally that Fisher was a genius of the first rank, who has a solid claim to being the most important applied mathematician of the 20th century. His work has a unique quality of daring mathematical synthesis combined with the utmost practicality. The stamp of his work is very much upon our field and shows no sign of fading. It is the stamp of a great thinker, and statistics—and science in general—is much in his debt.

REFERENCES

Section 1

SAVAGE, L. J. H. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist.* **4** 441–500. (Savage says that Fisher's work greatly influenced his seminal book on subjective Bayesianism. Fisher's great ideas are examined lovingly here, but not uncritically.)

YATES, F. and MATHER K. (1971). Ronald Aylmer Fisher. In *Collected Papers of R. A. Fisher* (K. Mather, ed.) **1** 23–52. Univ. Adelaide Press. (Reprinted from a 1963 Royal Statistical Society memoir. Gives a nontechnical assessment of Fisher's ideas, personality and attitudes toward science.)

BOX, J. F. (1978). *The Life of a Scientist*. Wiley, New York. (This is both a personal and an intellectual biography by Fisher's daughter, a scientist in her own right and also an historian of science, containing some unforgettable vignettes of precocious mathematical genius mixed with a difficulty in ordinary human interaction. The sparrow quote in Section 4 is put in context on page 130.)

Section 2

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 200–225. (Reprinted in the Mather collection, and also in the 1950 Wiley Fisher collection *Contributions to Mathematical Statistics*. This is my choice for the most important single paper in statistical theory. A competitor might be Fisher's 1922 Philosophical Society paper, but as Fisher himself points out in the Wiley collection, the 1925 paper is more compact and businesslike than was possible in 1922, and more sophisticated as well.)

EFRON B. (1995). The statistical century. *Royal Statistical Society News* **22** (5) 1–2. (This is mostly about the postwar boom in statistical methodology and uses a different statistical triangle than Figure 8.)

Section 3

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. Ser. A* **144** 285–307. (Concerns two situations when fully efficient estimation is possible in finite samples: one-parameter exponential families, where the MLE is a sufficient statistic, and location–scale families, where there are exhaustive ancillary statistics. Reprinted in the Mather and the Wiley collections.)

Section 4

EFRON, B. (1978). Controversies in the foundations of statistics. *Amer. Math. Monthly* **85** 231–246. (The Bayes–Frequentist–Fisherian argument in terms of what kinds of averages should the statistician take. Includes Fisher's famous circle example of ancillarity.)

EFRON, B. (1982). Maximum likelihood and decision theory. *Ann. Statist.* **10** 240–356. (Examines five questions concerning maximum likelihood estimation: What kind of theory is it? How is it used in practice? How does it look from a frequentistic decision-theory point of view? What are its principal virtues and defects? What improvements have been suggested by decision theory?)

CIFARELLI, D. and REGAZZINI, E. (1996). De Finetti's contribution to probability and statistics. *Statist. Sci.* **11** 253–282. [The second half of the quote in my Section 4, their Section 3.2.2, goes on to criticize the Neyman–Pearson school. De Finetti is less kind to Fisher in the discussion following Savage's (1976) article.]

Sections 5 and 6

DICICCIO, T. and EFRON, B. (1996). Bootstrap confidence intervals (with discussion). *Statist. Sci.* **11** 189–228. (Presents and discusses the cd4 data of Figure 2. The bootstrap confidence limits in Table 1 were obtained by the BC_a method.)

Section 7

REID, N. (1995). The roles of conditioning in inference. *Statist. Sci.* **10** 138–157. [This is a survey of the p^* formula, what I called the magic formula following Ghosh's terminology, and many other topics in conditional inference; see also the discussion (following the companion article) on pages 173–199, in particular McCullagh's commentary. Gives an extensive bibliography.]

EFRON, B. and HINKLEY, D. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65** 457–487. (Concerns ancillarity, approximate ancillarity and the assessment of accuracy for a MLE.)

Section 8

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. (Shows how approximate confidence intervals can be used to get good approximate confidence densities, even in complicated problems with a great many nuisance parameters.)

Section 9

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.* **37** 36–48. (The chronic hepatitis example is discussed in Section 10 of this bootstrap–jackknife survey article.)

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 99–138. (This paper and the ensuing discussion, occasionally rather heated, give a nice sense of Bayesian model selection in the Jeffreys tradition.)

KASS, R. and RAFTERY, A. (1995). Bayes factors. *J. Amer. Statist. Soc.* **90** 773–795. (This review of Bayesian model selection features five specific applications and an enormous bibliography.)

Section 10

EFRON, B. (1996). Empirical Bayes methods for combining likelihoods. *J. Amer. Statist. Assoc.* **91** 538–565.

Section 11

KASS, R. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91** 1343–1370. (Begins "Subjectivism has become the dominant philosophical tradition for Bayesian inference. Yet in practice, most Bayesian analyses are performed with so-called noninformative priors. . . .")

Section 12

LINDLEY, D. V. (1974). The future of statistics—a Bayesian 21st century. In *Proceedings of the Conference on Directions for Mathematical Statistics*. Univ. College, London.

SMITH, A. (1995). A conversation with Dennis Lindley. *Statist. Sci.* **10** 305–319. (This is a nice view of Bayesians and Bayesianism. The 2020 prediction is attributed to de Finetti.)

Comment

D. R. Cox

I very much enjoyed Professor Efron's eloquent and perceptive assessment of R. A. Fisher's contributions and of their current relevance. I am sure that Professor Efron is right to attach outstanding importance to Fisher's ideas.

As Professor Efron emphasizes, Fisher's ideas are so wide ranging that it is not feasible to cover them all in a single paper. The following outline notes are in supplementation of rather than in disagreement with the paper.

(1) Fisher stressed the need for different modes of attack on different types of inferential problem.

(2) While his formal ideas deal with fully parametric problems he gave the "exact" randomization test based on the procedure used in design. The status to him of such tests is not entirely clear. Did he regard them as reassurance for the faint of heart, timid about working assumptions of normality, or are they the preferred method of analysis to which normal theory based results are often a convenient approximation? Yates vigorously rejected the second interpretation. The more important point is probably that Fisher recognized that randomization indicated the appropriate analysis of variance, that is, appropriate estimate of error. This replaced the need for special ad hoc assumptions of a new linear model for each design.

(3) A key to understanding some of the distinctions between Fisherian and Neyman–Pearson approaches lies in Fisher's special notion of the meaning of the probability p of a "unique" event as set out, for example, in Fisher (1956, pages 31–36). There are two aspects, one that the individual belongs to an ensemble or population in a proportion p of which which the event holds. The other is that it is not possible to recognize the individual as lying in a subpopulation with a different proportion. Fisher considered, it seems to me correctly, that this enabled probability statements to be attached to an unknown parameter on the basis of a random sample, no other information being available, without invoking a prior distribution. The snag is that such distributions cannot be manipulated or combined by the ordinary laws of probability.

(4) The development of Fisher's ideas on Bayesian inference can be traced by comparing the polemical remarks at the start of *Design of Experiments* (Fisher, 1935) with the more measured comments in Fisher (1956).

(5) Fisher was of course an extremely powerful mathematician especially with distributional and combinatorial calculations. It helps us to understand his attitude to mathematical rigor to note the remarks of Mahalanobis (1938), who wrote:

> The explicit statement of a rigorous argument interested him but only on the important condition that such explicit demonstration of rigor was needed. Mechanical drill in the technique of rigorous statement was abhorrent to him, partly for its pedantry, and partly as an inhibition to the active use of the mind. He felt it was more important to think actively, even at the expense of occasional errors from which an alert intelligence would soon recover, than to proceed with perfect safety at a snail's pace along well known paths with the aid of a most perfectly designed mechanical crutch.

D. R. Cox is an Honorary Fellow, Nuffield College, Oxford OX1 1NF, United Kingdom (e-mail: david.cox@nuf.ox.ac.uk).