MATHEMATICAL PERSPECTIVES

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 50, Number 1, January 2013, Pages 129–146
S 0273-0979(2012)01374-5
Article electronically published on April 25, 2012

A 250-YEAR ARGUMENT: BELIEF, BEHAVIOR, AND THE BOOTSTRAP

BRADLEY EFRON

Abstract. The year 2013 marks the 250th anniversary of Bayes rule, one of the two fundamental inferential principles of mathematical statistics. The rule has been influential over the entire period—and controversial over most of it. Its reliance on prior beliefs has been challenged by frequentism, which focuses instead on the behavior of specific estimates and tests under repeated use. Twentieth-century statistics was overwhelmingly behavioristic, especially in applications, but the twenty-first century has seen a resurgence of Bayesianism. Some simple examples are used to show what's at stake in the argument. The bootstrap, a computer-intensive inference machine, helps connect Bayesian and frequentist practice, leading finally to an empirical Bayes example of collaboration between the two philosophies.

1. Introduction

Controversies, especially extended controversies, seem foreign to the precise discipline of mathematics. Mathematical statistics, nevertheless, is approaching the 250th anniversary of a fundamental dispute over the proper approach to scientific inference. I chose this dispute as the subject of the 85th Gibbs Lecture, delivered at the Joint Mathematics Meetings in January 2012, and what follows is essentially the text of that talk. The talk format emphasizes broad brushstrokes over rigor, with apologies necessary in advance for trampling nuance and detail. Some references (withheld until the end) provide a more careful review.

Received by the editors February 8, 2012, and, in revised form, February 10, 2012.
2010 *Mathematics Subject Classification.* Primary 97K70.
The author's work in supported in part by NIH grant 8R37 EB002784.

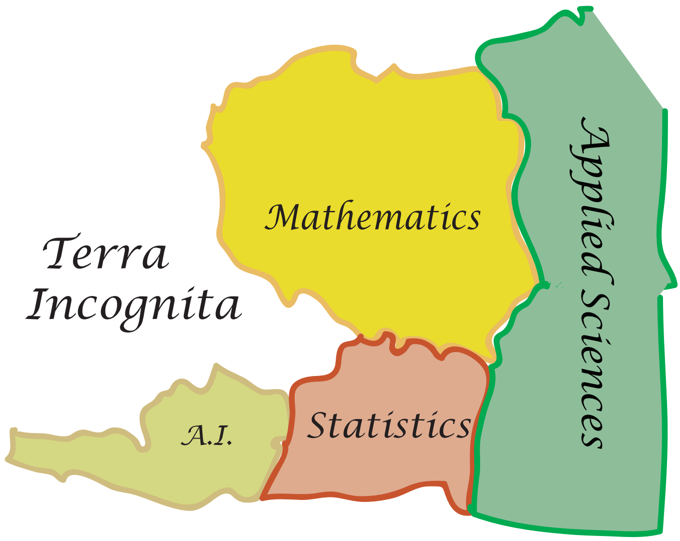

FIGURE 1. The greater world of mathematics and science (with thanks to Susan Holmes).

Figure 1 is supposed to be a schematic map of the greater mathematical world. Statistics shows up on it as a rugged frontier country, sharing a long border with the vast land of Applied Sciences lying to the east.

By and large, Statistics is a prosperous and happy country, but it is not a completely peaceful one. Two contending philosophical parties, the Bayesians and the frequentists, have been vying for supremacy over the past two-and-a-half centuries. The twentieth century was predominantly frequentist, especially in applications, but the twenty-first has seen a strong Bayesian revival (carried out partially in the name of Gibbs!).

Unlike most philosophical arguments, this one has important practical consequences. The two philosophies represent competing visions of how science progresses and how mathematical thinking assists in that progress.

My main goal here is to use very simple examples, of the type suitable for the after-dinner hour, to show what the argument is about. Not all is in dispute. Near the end of the talk I will give some hints of an emerging Bayesian/frequentist alliance, designed to deal with the enormous and complicated data sets modern scientific technology is producing. First though, I begin with a thumbnail sketch of Bayesian history and practice.

2. The physicist's twins

A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that she was going to have twin boys. One afternoon at the student union she suddenly asked me, "What are the chances my twins will be *identical* rather than *fraternal*"?

As an experienced statistical consultant I stalled for time, and asked if the doctor had told her anything else. “Yes, he said that one-third of twin births are *identical* and two-thirds are *fraternal*”.

Bayes would have died in vain if I didn’t use his rule to answer the question. We need to combine two pieces of partially contradictory evidence. Past experience favors *fraternal* according to the doctor, the prior odds ratio being

$$\frac{\Pr\{\text{identical}\}}{\Pr\{\text{fraternal}\}} = \frac{1/3}{2/3} = \frac{1}{2} \qquad \text{(prior experience).}$$

Current evidence observed from the sonogram, however, favors *identical*: identical twins are always the same sex while fraternals are equally likely to be the same or different sexes. In statistics terminology, the “likelihood ratio” of the current evidence is two-to-one in favor of *identical*,

$$\frac{\Pr\{\text{same sex}|\text{identical}\}}{\Pr\{\text{same sex}|\text{fraternal}\}} = \frac{1}{1/2} = 2 \qquad \text{(current evidence).}$$

(The gender, “boys” in this case, doesn’t affect the calculation.)

Bayes rule is a formula for combining the two pieces of information to get the posterior odds (or “updated beliefs”). In this situation the rule is very simple,

$$\begin{aligned} \text{Posterior odds} &= \text{Prior odds} \cdot \text{Likelihood ratio} \\ &= \frac{1}{2} \cdot 2 = 1; \end{aligned}$$

that is, equal odds for *identical* or *fraternal*. So I answered “50/50”, which greatly disappointed my friend who thought I was just guessing. (If I had said “three-sevenths to four-sevenths”, she would have considered me brilliant.)

The twins problem has a nice simple structure, with only two possible “states of nature”, *identical* or *fraternal*, and only two possible observations, *Same sex* or *Different*. This allows us to directly verify that 50/50 is the correct answer. The two-by-two table in Figure 2 shows the four possiblilites, *identical* or *fraternal* as rows, *Same sex* or *Different* as columns, with the four boxes labeled *a, b, c, d*.

We know two scientific facts: that box *b*, the upper right corner, has zero probability since identical twins are always the same sex; and that in the bottom row,

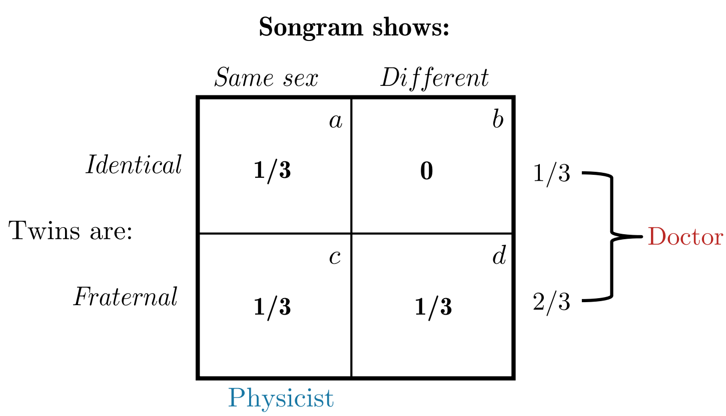

FIGURE 2. Analyzing the twins problem.

c and d have equal probabilities since the sexes of fraternal twins are independent of each other.

The doctor has also given us an important empirical fact: that the top row probabilities add up to 1/3 and the bottom row to 2/3. Putting all of this together gives the table of Figure 2, with zero probability in box b and 1/3 probability each in a, c, and d. The sonogram told the physicist that she was in the first column, *Same sex*, where the two possible states of nature have equal probability—so her odds were indeed 50/50. (I was tempted later to show her this table just to prove I wasn't guessing.)

3. Bayes rule

The twins example contains, in simplified form, all the elements of a general problem in Bayesian inference. We have:

- An unknown state of nature θ that we wish to learn more about. (θ equals *identical* or *fraternal* for the twins problem.)
- Prior beliefs about θ that can be expressed as a probability distribution $\pi(\theta)$. ($\pi(\theta)$ equals 1/3 or 2/3 for *identical* or *fraternal*.)
- An observation x that tells us something about θ. (From the sonogram.)
- And a probability model $f_\theta(x)$ that says how x is distributed for each possible value of θ. (As I described for the two-by-two table.)

The question is, having observed x, how should we update our "prior beliefs" $\pi(\theta)$ to "posterior beliefs" $\pi(\theta|x)$?

This brings us to Bayes rule, which will celebrate its 250th birthday next year. Thomas Bayes was a non-conformist English minister of substantial mathematical interest. (He would probably be a math professor these days.) His death in 1761 *was* almost in vain, but his friend Richard Price had Bayes rule, or theorem, published posthumously in the 1763 *Transactions of the Royal Society.* (Price thought the rule was a proof of the existence of God, an attitude not entirely absent from the current Bayesian literature.)

Bayes provided an elegant solution to the inference problem, that automates the construction I carried out for the two-by-two table: your posterior beliefs for θ having observed x are proportional to your prior beliefs times the likelihood of θ given x. As stated formulaically,

$$\text{Bayes rule:} \quad \pi(\theta|x) = c\pi(\theta) \cdot f_\theta(x).$$

Here c is just a proportionality constant that makes $\pi(\theta|x)$ integrate to one. The crucial elements are $\pi(\theta)$, the prior belief distribution, and the *likelihood* $f_\theta(x)$.

"Likelihood" has its technical meaning: the probability function $f_\theta(x)$ with the observed data x held fixed and the parameter θ varying—exactly the opposite of what we usually teach in our probability classes. As a simple example of a likelihood, suppose x follows a normal, or Gaussian, distribution centered at θ,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x)^2}.$$

Having observed say $x = 5$, the likelihood of θ would be a nice bell-shaped curve with its high point at 5.

The doctor provided us with a convincing prior distribution, presumably based upon a large amount of census data: the population proportion p of *identical* twins is 1/3. Bayes rule wouldn't be controversial if we always had genuine priors. Scientists, however, like to work on new questions, where prior experience is thin on the ground. The Bayesian/frequentist controversy *centers on the use of Bayes rule in the absence of genuine prior experience.*

For this we can thank Laplace, who seems to have discovered Bayes rule for himself in the early 1770s. Laplace followed Jacob Bernoulli in advocating the "Principle of Insufficient Reason". Applied to the twins example, (and in the absence of the doctor's information) the principle would say to take the population proportion p to be a priori uniformly distributed between 0 and 1, a "flat prior" for the unknown parameter.

Laplace's prestige was enough to carry the principle of insufficient reason through a century of scientific use. Venn in the 1870s began a counterattack, carried on in the 1900s by Keynes and R.A. Fisher, who pointed out that the principle was inconsistent in the sense that it produced different answers if applied, say, to the square root of p rather than p itself.

This didn't end the desire to use Bayes rule in situations without genuine prior experience. Harold Jeffreys, arguably the world's leading geophysicist at the time, devised an improved principle that was invariant under transformations. (More on this a little later.) For the twins problem, his rule would take the prior density for p, the population proportion of *identical* twins, not to be flat but rather U-shaped, going up sharply near zero and one,

$$\pi(p) = cp^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}.$$

"Objective Bayes" is the contemporary name for Bayesian analysis carried out in the Laplace–Jeffreys manner.

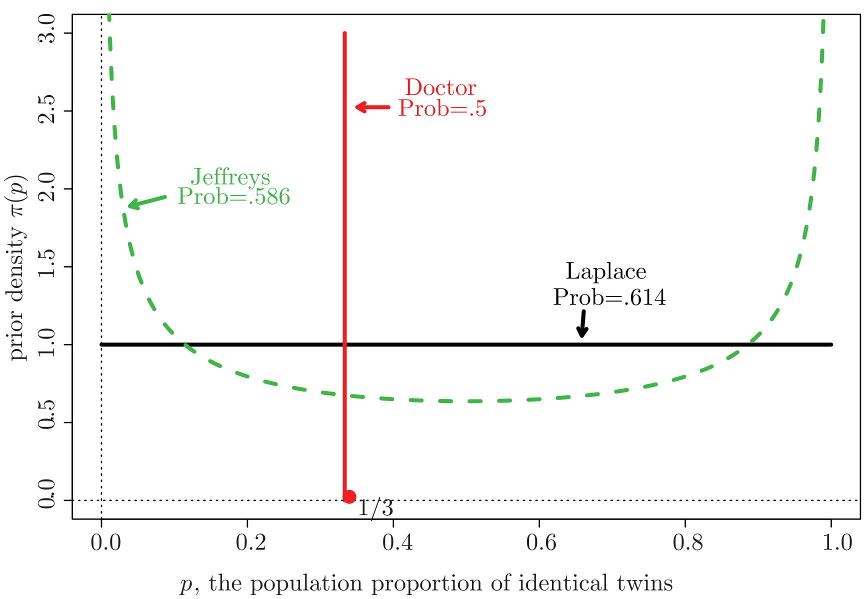

FIGURE 3. Three possible prior densities for p, the population proportion *identical*, and their predictions for the physicist's twins.

Figure 3 graphs three different prior distributions for p: the doctor's delta function at $p = 1/3$, Laplace's flat prior, and Jeffrey's U-shaped prior density. Of course different prior distributions produce different results. My answer to the physicist, that she had 50% chance of *identical* twins, changes to 58.6% with Jeffreys prior, and to a whopping 61.4% with a flat Laplace prior.

As I said earlier, there has recently been a strong Bayesian revival in scientific applications. I edit an applied statistics journal. Perhaps one quarter of the papers employs Bayes theorem, and most of these do *not* begin with genuine prior information. Jeffreys priors, and their many modern variations, are the rule rather than the exception. They represent an aggressive approach to mathematical modeling and statistical inference.

A large majority of working statisticians do not fully accept Jeffreys–Bayes procedures. This brings us to a more defensive approach to inference—frequentism, the currently dominant statistical philosophy.

4. Frequentism

Frequentism begins with three of the same four ingredients as Bayes theory: an unknown parameter or state of nature, θ, some observed data x, and a probability model $f_\theta(x)$. What is missing is $\pi(\theta)$, the prior beliefs. In place of $\pi(\theta)$, attention focuses on some statistical procedure that the statistician intends to use on the problem at hand. Here I will call it $t(x)$, perhaps an estimate or a confidence interval, or a test statistic or a prediction rule.

Inference is based on the behavior of $t(x)$ in repeated long-term use. For example, a prediction rule might be shown to be correct at least 90% of the time, no matter what the true θ happens to be. In this framework, *optimality*, finding the procedure $t(x)$ that has the best long-term behavior, becomes the central mathematical task. One might, for instance, look for the prediction rule with the smallest possible error rates. (Bayes theory has no need for optimality calculations since, within its own framework, the rule automatically provides ideal answers.) Optimality theory is where mathematics has played its greatest role in statistics.

The frequentist bandwagon really got rolling in the early 1900s. Ronald Fisher developed the maximum likelihood theory of optimal estimation, showing the best possible behavior for an estimate, and Jerzy Neyman did the same for confidence intervals and tests. Fisher's and Neyman's procedures were an almost perfect fit to the scientific needs and the computational limits of twentieth century science, casting Bayesianism into a shadow existence.

Figure 4 shows a small data set of the type familiar to Fisher and Neyman. Twenty-two students have each taken two tests, called "mechanics" and "vectors". Each of the points represents the two scores for one student, ranging from the winners at the upper right to the losers at lower left. We calculate the sample correlation coefficient between the two tests, which turns out to equal 0.498, and wonder how accurate this is.

What I previously called the data x is now all 22 points, while the statistic, or "method" $t(x)$ is the sample correlation coefficient. If the points had fallen exactly along a straight line with positive slope the sample correlation coefficient would be 1.00, in which case the mechanics score would be a perfect predictor for the vectors score, and vice versa (and they wouldn't have had to give two tests). The actual observed value, 0.498, suggests a moderate but not overwhelming

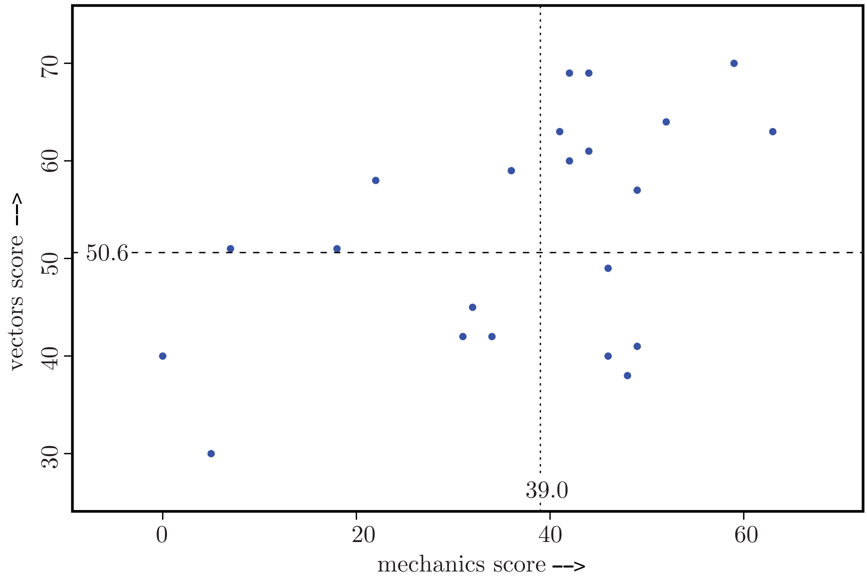

FIGURE 4. Scores of 22 students on two tests, "mechanics" and "vectors"; sample correlation coefficient is 0.498±??

predictive relationship. Twenty-two points is not a lot, and we might worry that the correlation would be much different if we tested more students.

A little bit of notation: $n = 22$ is the number of students; y_i is the data for the ith student, that is, his or her two scores; and the full data set y is the collection of all 22 y_i's. The parameter of interest, the unknown state of nature θ, is the true correlation—the correlation we would see if we had a much larger sample than 22, maybe even all possible students.

Now I've called the sample correlation coefficient 0.498 "$\hat{\theta}$". This is Fisher's notation, indicating that the statistic $\hat{\theta}$ is striving to estimate the true correlation θ. The "±??" after 0.498 says that we'd like some idea of how well $\hat{\theta}$ is likely to perform.

R.A. Fisher's first paper in 1915 derived the probability distribution for the correlation problem: what I previously called $f_\theta(x)$, now $f_\theta(\hat{\theta})$ with $\hat{\theta}$ taking the place of x. (It's a rather complicated hypergeometric function.) Much more importantly, between 1920 and 1935 he developed the theory of maximum likelihood estimation (MLE), and the optimality of the MLE. Speaking loosely, maximum likelihood is the best possible frequentist estimation method, in the sense that it minimizes the expected squared difference between $\hat{\theta}$ and the unknown θ, no matter what θ may be.

Fisher's 1915 calculations were carried out in the context of a bivariate normal distribution, that is, for a two-dimensional version of a bell-shaped curve, which I'll discuss a little later.

Despite pursuing quite similar scientific goals, the two founders of mathematical statistics, Fisher and Neyman, became bitter rivals during the 1930s, with not a good word to say for each other. Nevertheless, Neyman essentially completed

Fisher's program by developing optimal frequentist methods for testing and confidence intervals.

Neyman's 90% confidence interval for the student correlation coefficient is perhaps shockingly wide. It says that θ exists in

$$[0.164, 0.717],$$

with a 5% chance of missing on either end. Again speaking roughly, Neyman's interval is as short as possible in the absence of prior information concerning θ. The point estimate, $\hat{\theta} = 0.498$, looked precise, but the interval estimate shows how little information there actually was in our sample of 22 students.

Figure 5 is a picture of the Neyman construction, as physicists like to call it. The black curve is Fisher's probability distribution for $\hat{\theta}$ *if* the parameter θ equaled 0.164, the lower endpoint of the 90% confidence interval. Here 0.164 was chosen to put exactly 5% of the probability above the observed value $\hat{\theta} = 0.498$. Similarly the red curve, with parameter θ now the upper endpoint 0.717, puts 5% of the probability below 0.498.

A statistician who always follows Neyman's construction will in the long run cover the true value of θ 90% of the time, with 5% errors beyond each endpoint no matter what θ may be—the frequentist ideal—and with no prior beliefs required.

Maximum likelihood estimation and optimal testing and confidence intervals are used literally millions of times a year in serious scientific practice. Together they established the frequentist domination of statistical practice in the twentieth century.

We don't have any prior information for the student correlation problem, but that wouldn't stop Jeffreys or most modern Bayesians. Jeffreys prior, called objective or uninformative in the literature, takes the prior $\pi(\theta)$ to be proportional to one

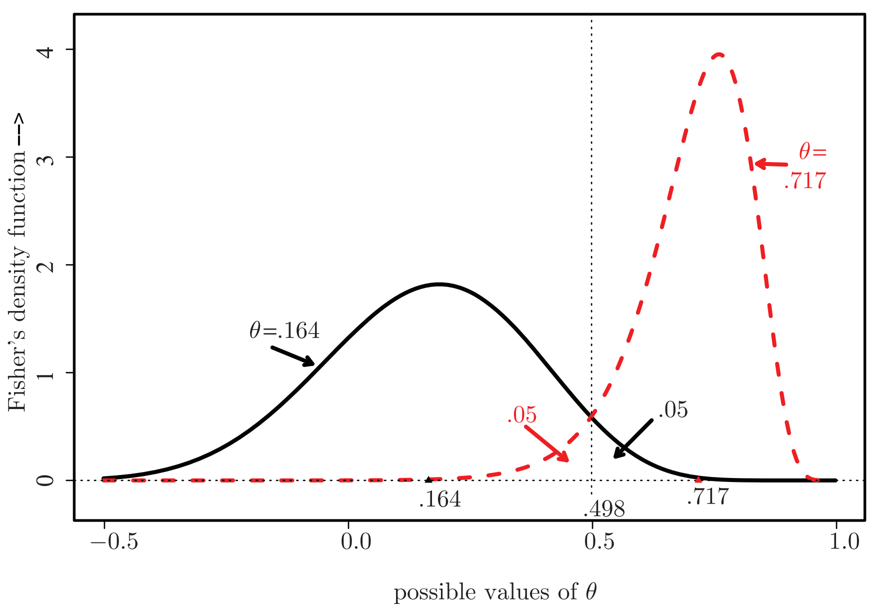

FIGURE 5. Neyman's 90% confidence interval for the student score correlation: $0.164 < \theta < 0.717$.

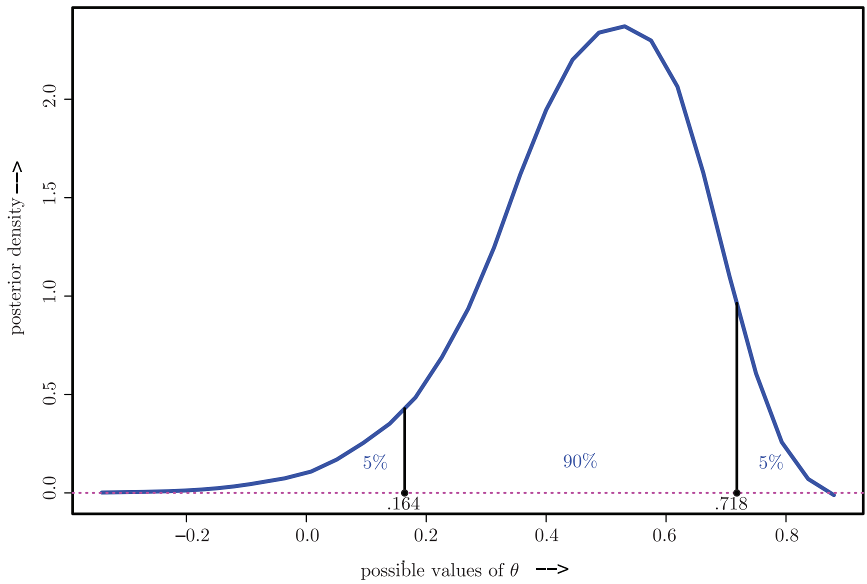

FIGURE 6. Jeffreys–Bayes posterior density $\pi(\theta|x)$ for the 22 students; 90% credible limits = $[0.164, 0.718]$; Neyman limits $[0.164, 0.717]$.

over $1 - \theta^2$,

$$\pi(\theta) = 1/(1 - \theta^2).$$

("Uninformative" is a positive adjective, indicating the hope that this choice of prior doesn't add any spurious information to the estimation of θ.)

Jeffreys' general formula depends, interestingly enough, on Fisher's information bound for the accuracy of maximum likelihood estimation, using it to generate priors invariant under monotonic transformation of the parameter.

I used Bayes rule, starting from Jeffreys prior, to estimate the posterior density of the student correlation coefficient given the 22 data points. The heavy curve in Figure 6 shows the posterior density $\pi(\theta|x)$, which is high near our point estimate 0.498, falling off asymmetrically, with a long tail toward the left, that is toward smaller values of θ. Five percent of the area under the curve lies to the left of 0.164 and 5% to the right of 0.718. These are the 90% Bayes limits for the unknown value of θ, almost exactly the same as Neyman's limits!

If we always had such nice agreement, peace would break out in the land of statistics. There is something special, however, about the correlation problem, which I'll get to soon.

I actually selected our 22 students randomly from a bigger data set of 88. Table 1 shows the sample correlation coefficient $\hat{\theta}$ as the sample size increased: at $n = 44$ the estimate jumped up from 0.498 to 0.663, coming down a bit to 0.621 at $n = 66$ and ending at 0.553 for all 88 students. The infinity row represents the unknown future, framed by the 90% Neyman interval based on all 88 students,

$$\theta \in [0.415, 0.662],$$

Table 1. More Students

n	$\hat{\theta}$
22	**.498**
44	.663
66	.621
88	.553
∞	[.415, .662]

now a good deal more precise than the interval [0.164, 0.717] based on only the original 22.

Statisticians usually do not have the luxury of peering into the future. Frequentism and Bayesianism are competing philosophies for extrapolating from what we can see to what the future might hold. That is what the 250-year argument is really about.

5. Nuisance parameters

Figure 7 represents the very first bivariate normal distribution, dating from 1886. It is due to Francis Galton, eccentric Victorian genius, early developer of fingerprint analysis and scientific weather prediction, and best-selling author of adventure travel books. Each of the 928 points shows an adult child's height along the y-axis and the parent's average height along the x-axis. The big red dot is at the two grand averages, about 68 inches each way in 1886. Somehow Galton realized that the points were distributed according to a two-dimensional correlated version of the bell-shaped curve. He was no mathematician but he had friends who were, and

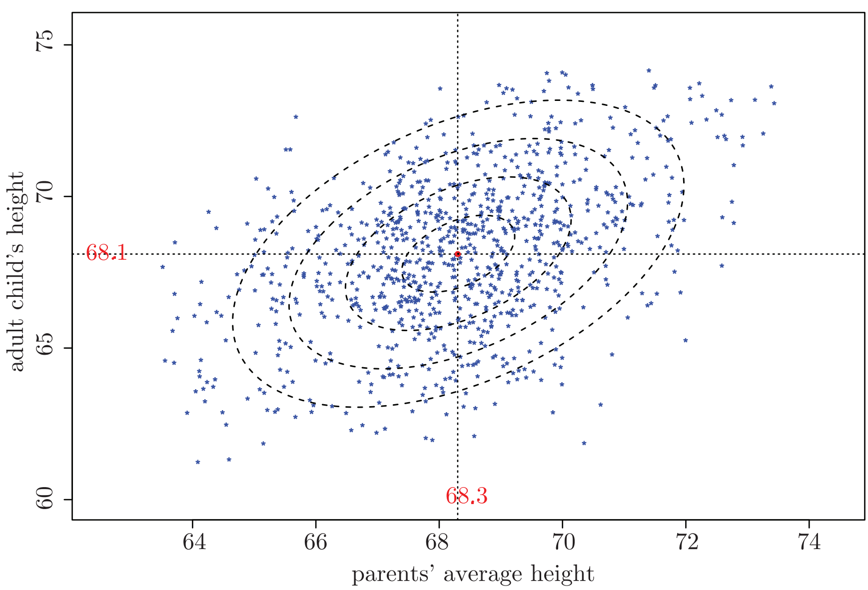

Figure 7. Galton's 1886 distribution of child's height vs parents'; ellipses are contours of best-fit bivariate normal density; the red dot is at bivariate average (68.3, 68.1).

they developed the formula for the bivariate normal density, which I'll discuss next. The ellipses show curves of equal density from the formula.

Galton was some kind of statistical savant. Besides the bivariate normal distribution, he used this picture to develop correlation (called by him originally "co-relation") and regression theory (called by him "regression to the mean": extremely tall parents have less extremely tall children, and conversely for shortness.)

Galton's formula for the probability density function of a bivariate normal random vector $y = (y_1, y_2)'$ is

$$f_{\mu,\Sigma}(y) = \frac{1}{2\pi} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}.$$

Here $\mu = (\mu_1, \mu_2)'$ is the two-dimensional *mean vector* while Σ is the 2-by-2 positive definite *covariance matrix.* (It describes the variabilities of y_i and y_2 as well as their correlation.) Standard notation for the distribution is

$$y \sim \mathcal{N}_2(\mu, \Sigma),$$

read "y is bivariate normal with mean μ and covariance $\boldsymbol{\Sigma}$". A perspective picture of the density function would show an esthetically pleasing bell-shaped mountain.

In Figure 7 I chose μ to match the red dot at the center and Σ to give the best-matching ellipses to the point cloud—in other words I used the maximum likelihood estimates of μ and $\boldsymbol{\Sigma}$. The main thing to note here is that a bivariate normal distribution has five free parameters—two for the mean vector μ and three for the symmetric matrix Σ—and all five will be unknown in typical applications.

For reasons having to do with relationships among the five parameters, the correlation problem turns out to be misleadingly easy. Here is a more difficult, and more typical, problem: suppose we are interested in the *eigenratio*, the ratio of the largest eigenvalue of the matrix Σ to the sum of the two eigenvalues,

$$\theta = \frac{\lambda_1}{\lambda_1 + \lambda_2} \qquad (\lambda_1 > \lambda_2 \text{ eigenvalues of } \boldsymbol{\Sigma}).$$

The MLE of Σ, $\hat{\Sigma}$, obtained from the 22 data points, gives the MLE

$$\hat{\theta} = 0.793 \pm ??,$$

where the question marks indicate that we want some assessment of how accurate $\hat{\theta}$ is for estimating the true value θ.

What is *not* true for the eigenratio is that the distribution of the quantity $\hat{\theta}$ we're interested in depends only on θ. This *was* true for the correlation, and effectively reduced all the calculations to one dimension. No matter how we try to reparameterize the five-parameter bivariate normal distribution, there will still be four *nuisance parameters* involved, in addition to the eigenratio, and they don't conveniently go away. Somehow they have to be taken into account before one can answer the $\pm$?? question.

Bayesian inference has a simple way of dealing with nuisance parameters: they are integrated out of the five-dimensional posterior distribution. However "simple" isn't necessarily "correct", and this can be a major point of practical disagreement between frequentist and Bayesian statisticians.

The heavy curve in Figure 8 is the Bayes posterior density for the eigenratio, starting from Jeffreys' five-dimensional uninformative prior and integrating out the four nuisance parameters. Dashed lines indicate the 90% Bayes posterior limits for the true eigenratio given the data for the 22 students. The red triangles are

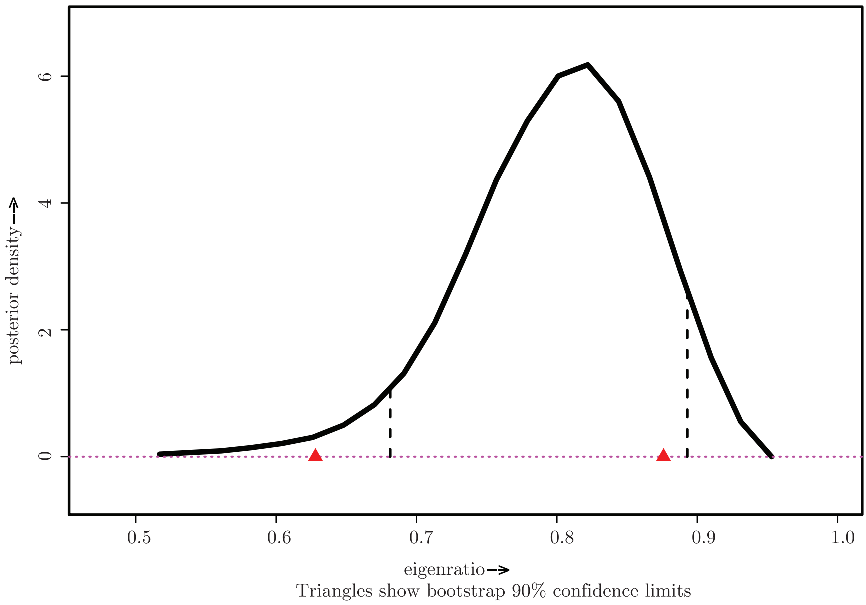

FIGURE 8. Posterior eigenratio density using Jeffreys prior for bivariate normal; 90% credible limits $[0.68, 0.89]$; bootstrap CI, triangles, $[0.63, 0.88]$.

frequentist 90% limits, obtained from a bootstrap calculation I will describe next. There is notable disagreement—the frequentist limits are shifted sharply downward.

Jeffreys prior, in fact, does not give frequentistically accurate confidence limits in this case, or in a majority of problems afflicted with nuisance parameters. Other, better, uninformative priors have been put forward, but for the kind of massive data analysis problems I'll discuss last, most Bayesians do not feel compelled to guarantee good frequentist performance.

6. The bootstrap and Gibbs sampling

I began this talk from a point in 1763, and so far have barely progressed past 1950. Since that time, modern scientific technology has changed the scope of the problems statisticians deal with and how they solve them. As I'll show last, data sets have inflated in size by factors of millions, often with thousands of questions to answer at once, swimming in an ocean of nuisance parameters. Statisticians have responded with computer-based automation of both frequentist and Bayesian technology.

The bootstrap is a frequentist machine that produces Neyman-like confidence intervals far beyond the point where theory fails us. Here is how it produced the eigenratio interval (the triangles in Figure 8). We begin with the data for the 22 students, along with a bivariate normal model for how they were generated,

$$y_i \sim \mathcal{N}_2(\mu, \Sigma), \qquad \text{independently for } i = 1, 2, \ldots, 22.$$

This gives maximum likelihood estimates $\hat{\mu}$ and $\hat{\Sigma}$ for the unknown mean and covariance of the normal distribution, as well as the eigenratio point estimate $\hat{\theta} = 0.793$.

In the bootstrap step we generate artificial data sets by sampling 22 times from the estimated bivariate normal distribution,

$$y_i^* \sim \mathcal{N}_2(\hat{\mu}, \hat{\Sigma}), \qquad \text{independently for } i = 1, 2, \ldots, 22.$$

Each such set provides a bootstrap version of the eigenratio, called $\hat{\theta}^*$ here.

I did the bootstrap calculations 10,000 times (a lot more than was actually needed), obtaining a nice histogram of the 10,000 bootstrap eigenratio estimates $\hat{\theta}^*$, as displayed in Figure 9. The mathematical theory behind the bootstrap shows how to use the histogram to get frequentistically accurate confidence intervals.

For example, the fact that a lot more than half of the $\hat{\theta}^*$'s exceed the original MLE of 0.793 (58% of them) indicates an upward bias in $\hat{\theta}$, which is corrected by moving the confidence limits downward. This is accomplished by reweighting the 10,000 $\hat{\theta}^*$'s, so that smaller values count more. (The dashed curve is the reweighting function.) The bootstrap confidence limits are the 5th and 95th percentiles of the reweighted $\hat{\theta}^*$'s.

The Bayesian world has also been automated. "Gibbs sampling" is a Markov Chain random walk procedure, named after Gibbs distribution in statistical physics. Given the prior and the data, Markov Chain Monte Carlo (MCMC) produces samples from an otherwise mathematically intractable posterior distribution $\pi(\theta|x)$. (The history of the idea has something to do with Los Alamos and the A-bomb.)

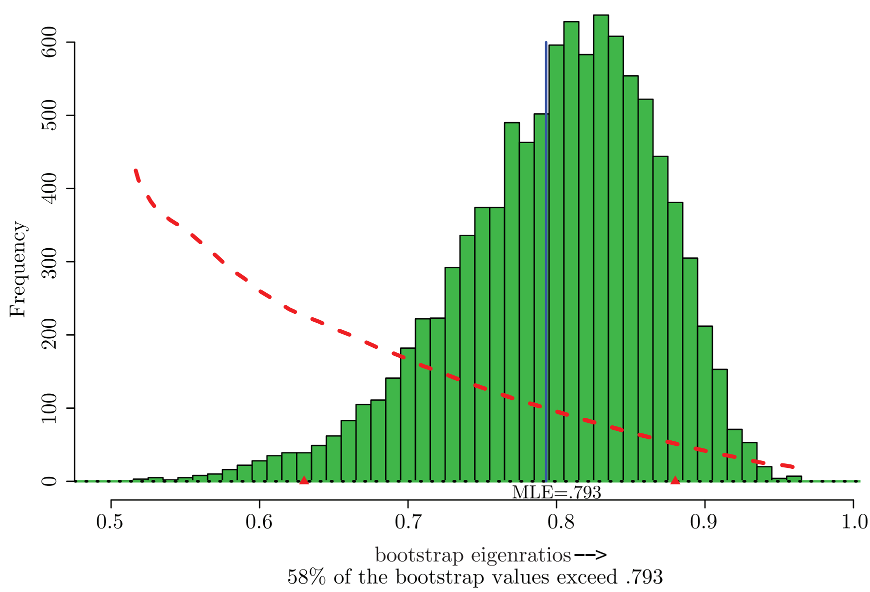

FIGURE 9. 10,000 bootstrap eigenratio values from the student score data (bivariate normal model); dashed line shows confidence weights.

MCMC theory is perfectly general, but in practice it favors the use of convenient uninformative priors of the Jeffreys style—which has a lot to do with their dominance in current Bayesian applications.

7. Empirical Bayes

I wanted to end with a big-data example, more typical of what statisticians are seeing these days. The data is from a prostate cancer study involving 102 men, 52 with prostate cancer and 50 healthy controls. Each man was measured on a panel of 6033 genes (using microarrays, the archetype of modern scientific high-throughput devices).

For each gene, a statistic x_i was calculated, the difference in means between the cancer patients and the healthy controls, which, suitably normalized, should be distributed according to a bell-shaped curve. For gene i, the curve would be centered at δ_i, the "true effect size",

$$x_i \sim \mathcal{N}(\delta_i, 1).$$

We can't directly observe δ_i, only its estimate x_i.

Presumably, if gene i doesn't have anything to do with prostate cancer, then δ_i will be near zero. Of course, the investigators were hoping to spot genes with big effects sizes, either positive or negative, as a clue to the genetic basis of prostate cancer.

The histogram in Figure 10 shows the 6033 effect size estimates x_i. The light dashed curve indicates what we would see if none of the genes had anything to do with prostate cancer, that is, if all the effect sizes were zero. Fortunately for the

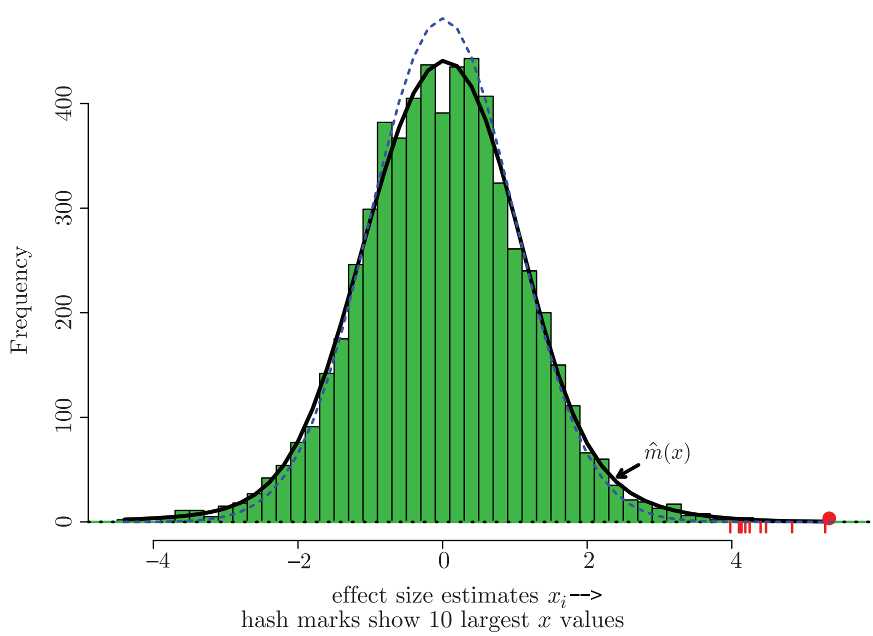

Figure 10. Prostate cancer study: difference estimates x_i comparing cancer patients with healthy controls, 6033 genes. Dashes indicate the 10 largest estimates.

investigators, that doesn't seem to be the case. A better fit to the histogram, called $\hat{m}(x)$, shows the heavier tails of the histogram, presumably reflecting genes with substantially big effect sizes.

Looking just at the right side, I've marked with little red dashes the 10 largest x_i's. These seem way too big to have zero effect size. In particular, the largest one of all, from gene 610, has $x_i = 5.29$, almost impossibly big if δ_{610} really equalled zero.

But we have to be careful here. With 6033 genes to consider at once, the largest observed values will almost certainly overstate their corresponding effect sizes. (Another example of Galton's regression to the mean effect.) Gene 610 has won a bigness contest with 6033 competitors. It's won for two reasons: it has a genuinely large effect size, and it's been lucky—the random noise in x_i has been positive rather than negative—or else it probably would not have won! The question is how to compensate for the competition effects and get honest estimates for the contest winners.

There's an intriguing Bayesian analysis for this situation. Considering any one gene, suppose its effect size δ has some prior density $\pi(\delta)$. We don't get to see δ, but rather x, which is δ plus some normal noise. If we know $\pi(\delta)$, we can use Bayes theorem to optimally estimate δ.

By definition, the *marginal density* of x is its density taking account of the prior randomness in δ and the normal noise,

$$m(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\delta)^2} \pi(\delta)\ d\delta.$$

Tweedie's formula is a neat expression for the Bayes posterior expectation of delta having observed x,

$$E\{\delta|x\} = x + \frac{d}{dx} \log m(x).$$

The trouble with applying Tweedie's formula to the prostate study is that without prior experience we don't know $\pi(\delta)$ or, therefore, $m(x)$. This is the kind of situation where frequentists rebel against using Bayesian methods.

There is, however, a nice compromise method available, that goes by the name "Empirical Bayes". If we draw a smooth curve through the green histogram, such as the heavy curve in Figure 10, we get a reasonable estimate $\hat{m}(x)$ of the marginal density $m(x)$. We can plug this into Tweedie's formula to estimate the Bayes posterior expectation of any one δ_i given its x_i,

$$\hat{E}\{\delta_i|x_i\} = x_i + \frac{d}{dx} \log \hat{m}(x)\bigg|_{x_i}.$$

At this point we've obtained a frequentist estimate of our Bayes expectation, without making any prior assumptions at all!

Figure 11 graphs the empirical Bayes estimation curve for the prostate study data. For gene 610 at the extreme right, its observed value $x = 5.29$ is reduced to an estimated effect size of 4.07 (a quantitative assessment of the regression to the mean effect). In a similar way, all of the x_i's are shrunk back toward zero, and it can be shown that doing so nicely compensates for the competition effects I was worried about earlier.

The curve has an interesting shape, with a flat spot between -2 and 2. This means that most of the genes, 93% of them, have effect size estimates near zero,

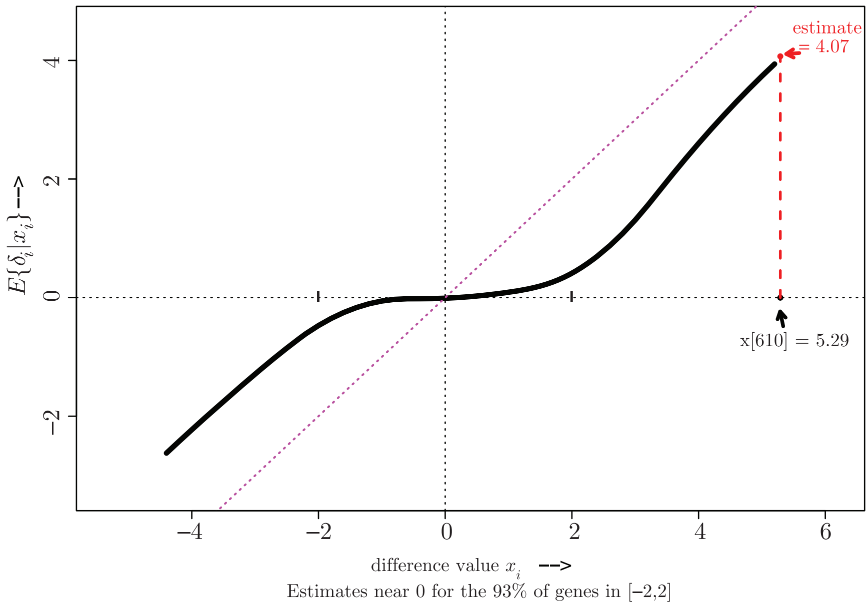

FIGURE 11. Empirical Bayes estimates of $E\{\delta|x\}$, the expected true difference δ_i given the observed difference x_i.

suggesting, sensibly, that most of the genes aren't involved in prostate cancer development.

Empirical Bayes is that Bayes/frequentist collaboration I referred to at the beginning of this talk—a hopeful sign for future statistical developments.

8. A score sheet

Table 2 is a score sheet for the frequentist/Bayesian argument, that you can use to decide which philosophical party you would join if you were an applied statistician.

(1) First and foremost, Bayesian practice is bound to prior beliefs, while frequentism focuses on behavior. The Bayesian requirement for a prior distribution, what I called $\pi(\theta)$, is a deal-breaker for frequentists, especially in the absence of genuine prior experience. On the other hand, frequentist analysis begins with the choice of a specific method, which strikes Bayesians as artificial and incoherent. Even optimal frequentist methods may be disparaged since the optmality refers to averages over hypothetical future data

TABLE 2. Score Sheet

Bayes	**Frequentist**
(1) Belief (prior)	(1) Behavior (method)
(2) Principled	(2) Opportunistic
(3) One distribution	(3) Many distributions (bootstrap?)
(4) Dynamic	(4) Static
(5) Individual (subjective)	(5) Community (objective)
(6) Aggressive	(6) Defensive

sets, different than the observed data x. This leads to a second major distinction:

(2) Bayesianism is a neat and fully principled philosophy, while frequentism is a grab-bag of opportunistic, individually optimal, methods. Philosophers of science usually come down strongly on the Bayesian side.

(3) Only one probability distribution is in play for Bayesians, the posterior distribution I called $\pi(\theta|x)$. Frequentists must struggle to balance behavior over a family of possible distributions, as illustrated with Neyman's construction for confidence intervals. Bayes procedures often have an alluringly simple justification, perhaps dangerously alluring according to frequentists. (Bootstrap methods are an attempt to reduce frequentism to a one-distribution theory. There are deeper Bayes/bootstrap connections than I have discussed here.)

(4) The simplicity of the Bayesian approach is especially appealing in dynamic contexts, where data arrives sequentially, and where updating one's beliefs is a natural practice. By contrast, frequentist methods can seem stiff and awkward in such situations.

(5) In the absence of genuine prior information, Bayesian methods are inherently subjective, though less so using objective-type priors of the Jeffreys sort. Bayesianism is very well suited to the individual scientist or small group, trying to make rapid progress on their subject of interest. Frequentism plays more to the wider scientific community, including skeptics as well as friends. Frequentism claims for itself the high ground of scientific objectivity, especially in contentious areas such as drug approval or faster-than-light neutrinos.

(6) My final criterion has to do with mathematical modeling. Bayesian theory requires an extra layer of modeling (for the prior distributions) and Bayesians tend to be aggressive math modelers. In fact, Bayesians tend to be more generally aggressive in their methodology. Another way to state the difference is that Bayesians aim for the best possible performance versus a single (presumably correct) prior distribution, while frequentists hope to due reasonably well no matter what the correct prior might be.

The two-party system can be upsetting to statistical consumers, but it has been a good thing for statistical researchers — doubling employment, and spurring innovation within and between the parties. These days there is less distance between Bayesians and frequentists, especially with the rise of objective Bayesianism, and we may even be heading toward a coalition government.

The two philosophies, Bayesian and frequentist, are more orthogonal than antithetical. And of course, practicing statisticians are free to use whichever methods seem better for the problem at hand — which is just what I do.

Meanwhile we can all get ready to wish Bayes rule a very happy 250th birthday next January.

About the author

Bradley Efron is professor of statistics and biostatistics at Stanford University. His research on statistical inference earned him the 2005 National Medal of Science.

References

[1] D. R. Bellhouse, *The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth*, Statist. Sci. **19** (2004), no. 1, 3–43; a biography of Bayes showing some of his other mathematical interests. MR2082145

[2] James Berger, *The case for objective Bayesian analysis*, Bayesian Anal. **1** (2006), no. 3, 385–402 (electronic); a spirited argument for Jeffreys-type Bayesian analysis. MR2221271

[3] Julian Champkin, *Francis Galton centenary*, Significance **8** (2011), no. 3, 121–121; three short articles on Galton's life and work including the child-parent regression example. MR2247161

[4] Bradley Efron, *Microarrays, empirical Bayes and the two-groups model*, Statist. Sci. **23** (2008), no. 1, 1–22; much more on the prostate cancer study. MR2431866

[5] ———, *Bayesian inference and the parametric bootstrap*, Submitted, Annals of Applied Statistics, 2012; connections between bootstrap and Bayes computation, the weighting curve of Figure 9, and Fisher's correlation formula and the student score example.

[6] ———, *Tweedie's formula and selection bias*, J. Amer. Statist. Assoc. **106** (2011), no. 496, 1602–1614; discussion of the formula behind Figure 11.

[7] Robert E. Kass and Larry Wasserman, *The selection of prior distributions by formal rules*, J. Amer. Statist. Assoc. **91** (1996), no. 435, 1343–1370; thorough discussion of Jeffreys priors in their original and modern forms.

[8] Erich L. Lehmann and Joseph P. Romano, *Testing Statistical Hypotheses*, 3rd ed., Springer Texts in Statistics, Springer, New York, 2005; Section 3.5 discusses Neyman's construction. MR2135927 (2006m:62005)

[9] Kantilal Varichand Mardia, John T. Kent, and John M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979; Table 1.2.1 gives the student score data. MR560319 (81h:62003)

Department of Statistics, 390 Serra Mall, Stanford, California 94305-4065
E-mail address: `brad@stat.stanford.edu`