BAYESIANS, FREQUENTISTS, AND PHYSICISTS

BRADLEY EFRON

Phystat2003 brought statisticians together with particle physicists, astrophysicists, and cosmologists. This paper, which is taken from the text of the keynote address, concerns the uneasy relationship between Bayesian and frequentist statistics, with particular attention to the "neutrino problem": how to set confidence limits for a parameter known to be non-negative. Model selection, an objective Bayes technique, gives a different answer than the classic Neyman confidence construction.

1. Introduction

Ten years ago I gave a talk entitled “Astronomy and biostatistics, the odd couple”. It emphasized the mostly unconscious convergence of methods in the two fields, arising from a shared need to account for biased sampling – astronomers because they are stuck on earth and statisticians because humans are such lousy experimental animals, tending to go missing just when you most need their data.

Of course astronomy historically is the most statistical branch of the physics tree. I never expected to attend a conference bringing particle physicists and statisticians under the same roof. The happy existence of phystat2003 reflects the determination of modern physicists to pursue nature in its most subtle manifestations, where the certainties of mass experimentation give way to small numbers of events observed with great difficulty; in short where statistical inference becomes important.

It is hard to imagine “phystat1903”. Maybe phystat1803 is slightly more conceivable with Laplace and Gauss arguing the virtues of Bayesian versus frequentist inference. The Bayesian-frequentist argument is certainly a long-lived one, even by the standards of philosophy. It reflects, I believe, two quite different attitudes toward the scientific process: the cautious frequentist desire for objectivity and consensus, versus the individual scientist trying aggressively to make the best sense of past data and the best choice for future direction.

Statistics concerns the efficient accumulation of knowledge – how a scientist learns from experience – so it is not surprising that there is more than one philosophical approach to such a broad problem. I will tread lightly on philosophical matters here. Mainly I want to show how Bayesian and frequentist ideas interact, in particular concerning confidence intervals and the “neutrino problem”, where both methodologies show virtues and defects.

2. Why Not Bayes?

Bayesianism was the first statistical philosophy, and remains the simplest and most attractive from the point of view of intellectual neatness. Here is a true-life story illustrating Bayesian virtues. A physicist friend of mine and her husband found out, thanks to a sonogram, that they were going to be the parents of twin boys. The doctor told them that one-third of twin pairs are identical, the other two-thirds fraternal, but she wondered if there was a more exact estimate for *her* twins being identical.

This is exactly the problem Bayes solved in 1763. To use standard language, the prior odds of Identical are $(1/3)/(2/3) = 1/2$, while the likelihood ratio (the ratio of probabilities of observing "Both Boys") equals about

$$\frac{P\{\text{Both Boys}|\text{Identical}\}}{P\{\text{Both Boys}|\text{Fraternal}\}} = \frac{1/2}{1/4} = 2.$$

Then Bayes' rule says that the *a posteriori* odds of Identical to Fraternal equal the prior odds times the likelihood ratio,

$$\frac{P\{\text{Identical}|\text{Both Boys}\}}{P\{\text{Fraternal}|\text{Both Boys}\}} = \frac{1}{2} \times 2 = 1.$$

In other words my physicist friend had a 50-50 chance for Identical. (Later she told me that the boys turned out to be "very non-identical".)

Bayes' rule is satisfying, convincing, and fun to use. But using Baye's rule does not make one a Bayesian; *always* using it does, and that's where difficulties begin. "Expert Opinion", which (correctly) gave us the prior odds of one-third to two-thirds for the Twins question, doesn't exist for most genuine scientific problems, or if it does exist may be contentious or just plain wrong. Even the likelihood ratio, which doesn't depend on prior opinions, may be impossible to compute.

Scientists are drawn to Bayesian thinking and language even when they can't use Bayes' theorem. In 2002 a heroic bout of radio telescopy detected subtle polarization in the cosmic microwave background, as predicted by the Big Bang theory. Michael Turner, a leading cosmologist, commented "If you had any doubts that this radiation is from the Big Bang, this should quash them." Even leaving aside the almost theological question of prior odds on the Big Bang theory, this is a case where the denominator of the likelihood ratio,

$$\frac{P\{\text{Polarized}|\text{Big Bang}\}}{P\{\text{Polarized}|\text{Any Other Theory}\}}$$

seems especially obscure. Dr. Turner spoke as a good scientist but a questionable Bayesian.

The 20th century saw the development of a persuasive frequentist theory of statistical inference that continues to dominate scientific practice. (An "objective" Bayesian counter-reformation is stirring, discussed later.) Frequentist statistics does away with the need for prior opinions. This gives frequentism a legitimate claim to objectivity, a considerable virtue in a world with competing scientific teams working at great distances from each other.

Here is a classic frequentist result: data points $x_1, x_2, \ldots, x_n$ are independently observed from a normal distribution with mean μ and variance 1,

$$x_1, x_2, \ldots, x_n \overset{\text{ind}}{\sim} N(\mu, 1), \tag{1}$$

and we want to say something about the unknown value of μ. The *standard 90% confidence interval* for μ is

$$\mu \in [\bar{x} - 1.645/\sqrt{n},\ \bar{x} + 1.645/\sqrt{n}\ ], \tag{2}$$

with $\bar{x}$ the mean $\Sigma x_i/n$. Interval (2) contains the unknown μ with 90% probability no matter what μ might be, which is its crucial frequentist property.

Formula (2) and its generalizations are familiar friends, appearing literally millions of times per year in the scientific literature. Nevertheless, there are limitations to the frequentist viewpoint that become more problematical as one moves away from simple situations like (1). One doesn't have to move very far away for difficulties to emerge, as the neutrino problem will show.

3. Accuracy and Correctness

One criticism of frequentist statistics is that it is incomplete: by itself it does not necessarily produce a unique solution to a given inferential problem. In situation (1) for instance the interval

$$\mu \in [\bar{x} - 1.96/\sqrt{n},\ \bar{x} + 1.44/\sqrt{n}\ ]$$

also contains μ with 90% probability. The standard interval (2) is preferred because of its symmetry, dividing its 10% noncoverage probability into 5% errors on either side, but symmetry is not part of the frequentist philosophy.

R.A. Fisher, the founder of modern statistical theory, and the person most responsible for demoting Bayesianism to the back burner, distrusted Jerzy Neyman's confidence interval methodology. He felt that Neyman intervals could be *accurate* (that is, have the claimed coverage probability) without being *correct.*

Here is a simple example of Fisherian "incorrectness". Suppose that in situation (1) the sample size n is either 25 or 100, the choice being determined by the independent flip of a fair coin. It is easy to compute that in this case

$$\bar{x} \pm .262 \tag{3}$$

is a 90% confidence interval for μ. However it is always incorrect: if $n = 25$ then interval (2) is $\bar{x} \pm .329$, wider than (3), while $n = 100$ gives the narrower interval $\bar{x} \pm .164$.

Fisher called the random variable n an *ancillary statistic*, a quantity conveying no direct information for μ, but whose value determines the accuracy with which μ can be estimated.

With typical ingenuity Fisher showed that ancillaries can pop up quite unexpectedly, even in innocuous-looking situations.

Suppose we independently observe $x_1, x_2, \dots x_{10}$ from a Cauchy distribution with unknown center μ, so each x_i has density

$$f_\mu(x) = \pi^{-1}/[1 + (x - \mu)^2].$$

The maximum likelihood estimate $\widehat{\mu}$ is, to a quite good approximation, normally distributed about μ,

$$\widehat{\mu} \mathrel{\dot\sim} N(\mu, \sigma^2) \quad \text{with} \quad \sigma = 0.447, \tag{4}$$

but the obvious 90% confidence interval $\mu \in \widehat{\mu} \pm 1.645 \cdot 0.447$, which is almost perfectly accurate, runs into the same correctness problems as (3).

Here the ancillary is the second derivative of the log likelihood function $\ell(\mu) = \log \prod_{i=1}^{10} f_\mu(x_i)$ evaluated at $\widehat{\mu}$,

$$\ddot{\ell} = \frac{\partial^2}{\partial \mu^2} \ell(\mu)|_{\mu = \hat{\mu}}.$$

The conditional distribution of $\widehat{\mu}$ given $\ddot{\ell}$ is approximately

$$\widehat{\mu} | \ddot{\ell} \sim N(\mu, (-\ddot{\ell})^{-1}),$$

so the magnitude of $-\ddot{\ell}$ determines the accuracy of $\widehat{\mu}$ for estimating μ. The correct interval is

$$\widehat{\mu} \pm 1.645 \cdot (-\ddot{\ell})^{-\frac{1}{2}}. \tag{5}$$

Figure 1 shows log likelihood functions for two Cauchy samples each of size 10, both of which have $\widehat{\mu} = 5.00$. However the "good sample" has its likelihood declining much more rapidly on either side of $\widehat{\mu}$, resulting in a shorter interval (5),

$$\text{Good Sample}: \ 5.00 \pm .35; \quad \text{Bad Sample}: \ 5.00 \pm .65. \tag{6}$$

(The difference is the closer spacing of the good samples' observations around $\widehat{\mu} = 5.00$.)

Both (5) and the unconditional interval $\widehat{\mu} \pm 1.645 \cdot 0.447$ are accurate, i.e. have 90% frequentist coverage, but (5) more correctly reflects the information available in the sample at hand. This problem doesn't affect interval (2) since $\bar{x}$ is a sufficient statistic in the normal case (1), but an argument can be made that Cauchy-type situations are actually more common.

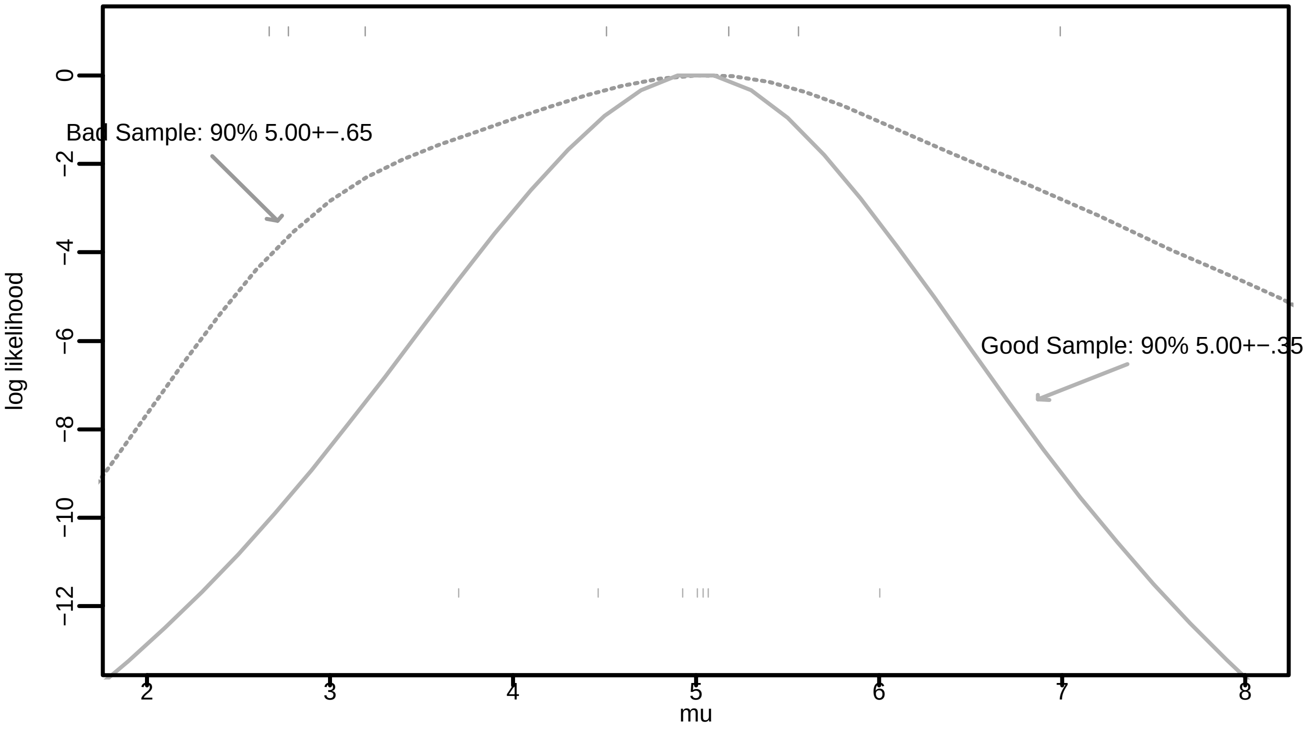

Figure 1:*Log likelihood functions for two Cauchy samples of size* $n = 10$*, each with* $\widehat{\mu} = 5.00$*;* μ *can be determined much more accurately in the "good sample", which has its likelihood decreasing more rapidly on either side of* $\widehat{\mu}$.

It's worth noting that the simplest Bayesian analysis, beginning with a flat prior for μ (an improper uniform distribution over the entire real line) automatically gives results like (6). This is a reminder that Bayesian properties are important to consider even if are eventually intends a frequentist analysis.

4. The Neutrino Problem

Adding a non-negativity constraint,

$$\mu \geq 0 \tag{7}$$

to the normal sampling assumptions (1) brings us to what might be called "the neutrino problem" since it arises pertinaciously in experiments assessing the mass of the neutrino. Here we will state the problem as setting a 95% upper limit for μ under model (1) and (7). There is no loss of generality taking $n = 1$ in (1), so we can express the data simply as a single observation of

$$x \sim N(\mu, 1). \tag{8}$$

The standard one-sided interval for μ in this case is

$$\mu \leq x + 1.645. \tag{9}$$

The standard interval in perfectly accurate, covering μ exactly 95% of the time no matter what its value might be. However it seems incorrect, especially to physicists, since if x is less than -1.645 interval (9) is empty of values satisfying (7). Exactly this case arose in actual neutrino experimentation, see Mandelkern (2002).

Various alternative bounds have been suggested, the best-known being the Feldman-Cousins bounds (1998), classic Neyman confidence limits based on likelihood ratio statistics. Figure 2 displays four possibilities: the standard bound $x + 1.645$; Feldman-Cousins; the Bayes upper 95% *a posteriori* value beginning with a uniform prior density for μ on $[0, \infty)$; and a bound based on *model selection*, as discussed below, an "objective Bayesian" construction that takes seriously the possibility of μ equaling zero. If a good bound is a small one then Model Selection is the clear winner, the Uniform$[0, \infty)$ Bayesian bound is worst, while Feldman-Cousins and the standard bound are intermediate, differing only for $x < 0$.But of course there is more to the story.

Which upper bound is right? The Bayesian answer would be "surely that depends upon our prior knowledge concerning μ. If there is nothing scientifically special about $\mu = 0$ except as the lower endpoint of the allowable μ domain, then a case can be made for the Uniform$[0, \infty)$ prior approach. In some situations though there may be strong scientific grounds for suspecting this μ equals zero, or at least is very close to zero compared to the standard deviation 1 of the observation $x \sim N(\mu, 1)$. Then an analysis should reflect those beliefs."

"Model Selection" refers to the last situation. We assume that there are two possible models for the parameter μ in (8),

$$\begin{array}{ll} \textit{Small Model} & \mathcal{M}_0 : \mu = 0 \\ \textit{Big Model} & \mathcal{M}_1 : \mu > 0. \end{array} \tag{10}$$

Having observed $x \sim N(\mu, 1)$ we wish to set a 95% upper bound for μ, including in our calculations the possibility that $\mu = 0$. Because the two models are of different dimensions, a hallmark of the general model selection problem, this is a more intricate task than setting confidence bounds within a single model.

A full Bayesian analysis requires prior probabilities on the two models in (10), and also

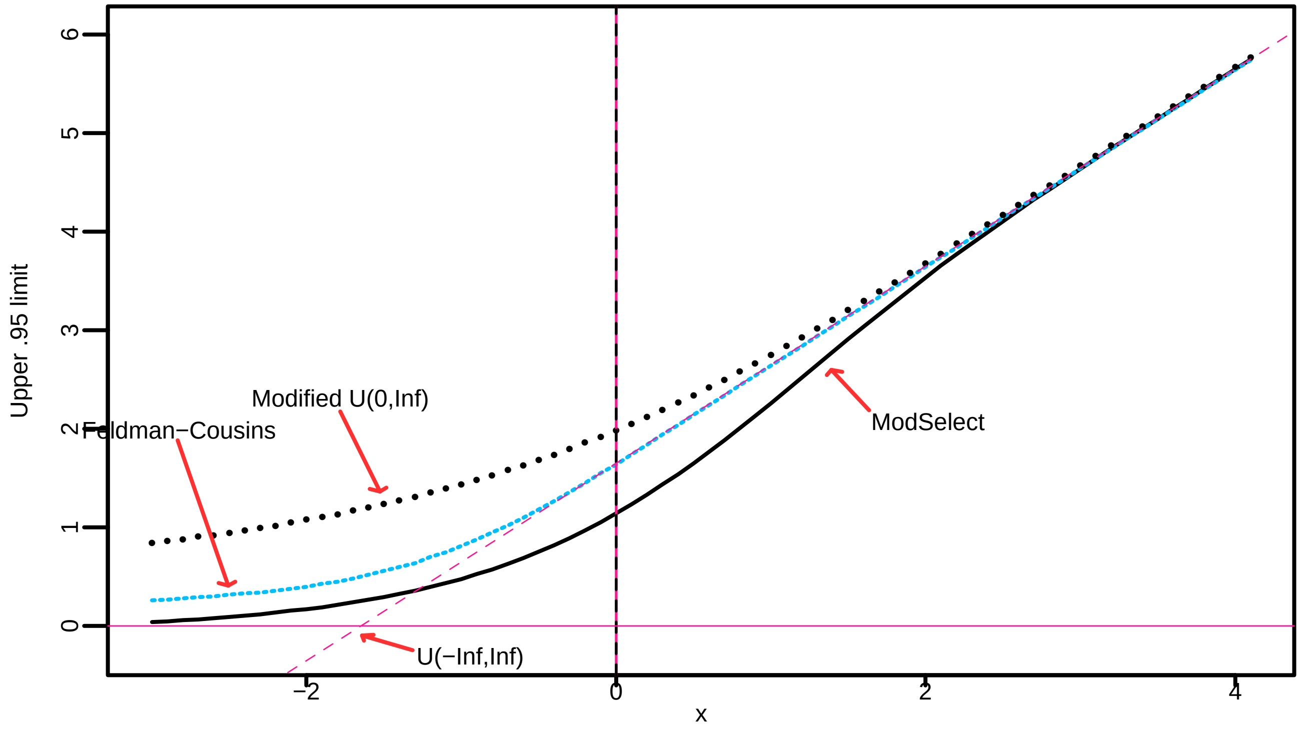

Figure 2:*Four possible 95% upper bounds for μ having observed x in model (7,8): standard bound $x + 1.645$; Feldman-Cousins; Bayesian bound versus uniform prior on $[0, \infty)$; Model Selection bound.*

a prior density on μ when $\mathcal{M}_1$ applies,

$$\textit{A priori} \begin{cases} \mathcal{M}_0 \text{ true with probability } \pi_0 \\ \mathcal{M}_1 \text{ true with probability } \pi_1 = 1 - \pi_0 \end{cases} \tag{11}$$

and

$$\mu \sim g(\cdot) \quad \text{if} \quad \mathcal{M}_1 \quad \text{true,} \tag{12}$$

where $g(\mu)$ is some prior density on $(0, \infty)$.

The Bayes 95% upper limit for μ having observed $x \sim N(\mu, 1)$ from prior situation (11, 12) depends on an integral of the sampling density $\varphi_\mu(x) = (2\pi)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\mu)^2\}$,

$$\varphi_{(1)}(x) \equiv \int_0^\infty \varphi_\mu(x) g(\mu) d\mu,$$

as well as $\varphi_0(x) = (2\pi)^{-\frac{1}{2}}\exp\{-x^2/2\}$. Bayes' rule then provides the *a posteriori* probability of $\mathcal{M}_1$ given x,

$$\text{Prob}\{\mathcal{M}_1|x\} = \pi_1\varphi_{(1)}(x)/[\pi_0\varphi_0(x) + \pi_1\varphi_{(1)}(x)], \tag{13}$$

and the probability that μ exceeds some positive value "c" given $\mathcal{M}_1$ and x,

$$\text{Prob}\{\mu > c|\mathcal{M}_1, x\} = \int_c^\infty \varphi_\mu(x)g(\mu)d\mu/\varphi_{(1)}(x). \tag{14}$$

Since μ can exceed c only if $\mathcal{M}_1$ is true, the *a posteriori* probability of $\mu > c$ is

$$\text{Prob}\{\mu > c|x\} = (13) \cdot (14). \tag{15}$$

The value of c that makes (15) equal is the Model Selection upper 95% bound.

At this point the non-Bayesian may quail at making prior selections of π_0 and $g(\mu)$. "Objective Bayes", currently the most popular form of Bayesian statistics, attempts to alleviate such fears by restricting attention to priors that enjoy reasonable frequentist properties.

First consider choosing $g(\mu)$ in (12). Without the negativity constraint (1) we might very well take $g(\mu) \equiv 1$, which yields the standard interval (9) as the one-sided Bayes 95% interval. This suggests using

$$g(\mu) = 1 \quad \text{on} \quad (0, \infty) \tag{16}$$

in (12), the prior that gave the dotted curve in Figure 2. (Another choice is mentioned below.) Now $\varphi_{(1)}(x)$ equals the standard normal cumulative distribution function

$$\Phi(x) = \int_{-\infty}^x \varphi_0(x)dx,$$

so (13) becomes

$$\text{Prob}\{\mathcal{M}_1|x\} = \pi_1\Phi(x)/[\pi_0\varphi_0(x) + \pi_1\Phi(x)]. \tag{17}$$

We can complete the choice of objective prior by selecting π_0, and $\pi_1 = 1 - \pi_0$, to make

$$\text{Prob}\{\mathcal{M}_1|X_0\} = \frac{1}{2} \tag{18}$$

at some "break-even point" x_0 that is comfortable for frequentists. Efron and Gous (2001) argue that

$$x_0 = 1.28 \tag{19}$$

the 90th percentile of x under $\mathcal{M}_0$, is a choice commensurate with standard Fisherian hypothesis testing. The "Model Select" curve in Figure 2 is formula (15) evaluated for the prior choices (16-19). [Note: solving $P\{\mathcal{M}_1|x_0\} = \frac{1}{2}$ in (17) determines π_0 and $\pi_1 = 1 - \pi_0$, so that (15) can be evaluated. However π_0 cannot be interpreted as the actual prior probability of $\mathcal{M}_0$; we would get a different π_0, but the same values of (15), if (16) were changed to say $g(\mu) = 2$ on $(0, \infty)$.]

Figure 3 graphs Prob$\{\mathcal{M}_0|x\}$, the *a posteriori* probability that $\mu = 0$. The values roughly agree with standard frequentist hypothesis testing: observing x equal the 95th $\mathcal{M}_0$ percentile 1.645 gives mild evidence against $\mathcal{M}_0$, Prob$\{\mathcal{M}_0|x\} = 0.36$; x equal the 99th percentile gives strong evidence, etc. Observing $x = 0$ gives 80% probability that $\mu = 0$. This leaves only 20% for $\mu > 0$, and leads to the notably small 95% upper bound.

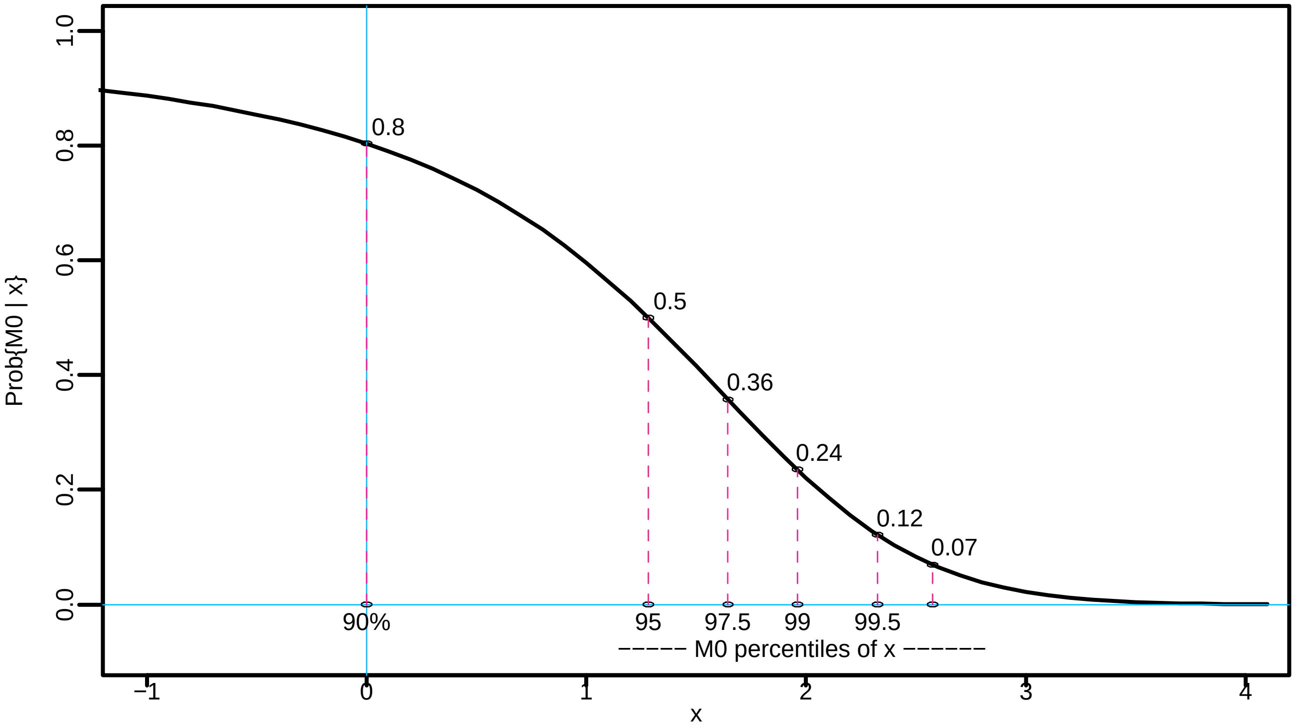

Figure 3: *A posteriori probability that* $\mu = 0$ *given* x *for Model Selection specifications (16)-(19); observing* $x = 1.645$ *gives 36% probability that* $\mu = 0$*, etc.*

From a classical point of view Model Selection looks like an ungainly combination of hypothesis testing and estimation. We used Bayes' theorem to handle the problem, but the resulting method is not subjective in the sense of employing specific prior knowledge (or guesses) concerning neutrinos. What it does employ is a qualitative belief that values of μ at or near 0 deserve increased weight. "Near" means near compared to the variance 1 of the observation $x \sim N(\mu, 1)$. Mandelkern's discussion suggests that this was and is the case for the actual electron neutrino problem.

It is easy to criticize Bayesian Model Selection methods. Even trying to be objective, we still have had to make an uncomfortable number of prior specifications. Alternative specifications are certainly possible. Choosing

$$g(\mu) = 1 + 3e^{-3\mu} \quad (\text{for } \mu > 0),$$

instead of $g(\mu) = 1$, makes the dotted curve in Figure 2 look much more like the standard bound $x + 1.645$ when $x \geq 0$. However this change doesn't have much effect on the Model Select curve.

Moving the break-even point x_0 closer to zero, which effectively puts less prior probability on $\mu = 0$, changes the Model Selection upper 95% value more dramatically. Taking $x_0 = 0.50$ instead of 1.28 makes the Model Selection upper 95% bounds agree rather nicely with the Feldman-Cousins curve, as seen in Figure 4.

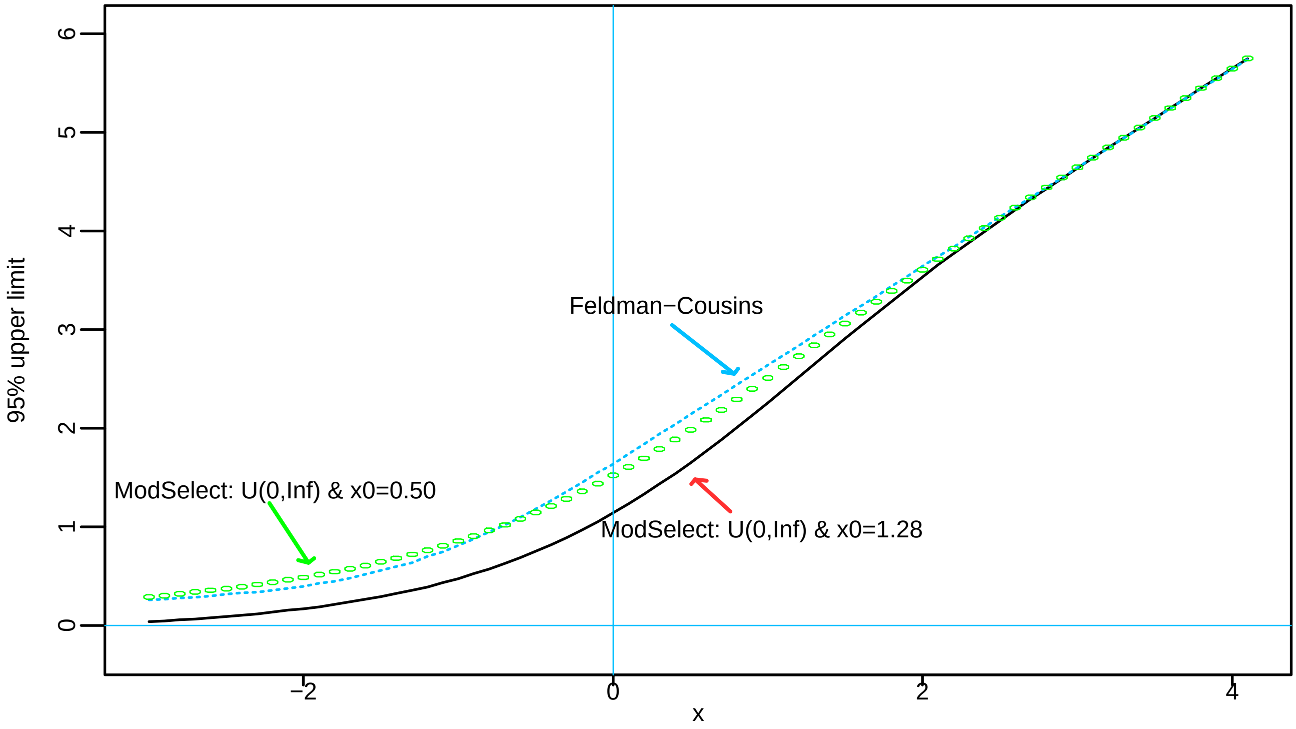

Figure 4: *Moving the break-even point x_0, (17), closer to zero makes the Model Selection upper 95% point nearly match the Feldman-Cousins bound.*

At this point an optimist could say that we have the best of all possible statistical worlds, with close agreement between the Bayesian Model Selection and Neyman frequentist approaches. However the two methods deliver there 95% upper bounds quite differently. At $x = -0.50$, roughly the neutrino result obtained by the "Troisk" group in (1999) according to Mandelkern, both bounds equal about 1.15; however Model Selection (with $x_0 = 0.50$)

implies a 69% probability that μ equals or is quite near 0. If believed, this might strongly influence subsequent detection strategies.

5. Multidimensional Problems and Stein's Paradox

The neutrino problem is one-dimensional in the sense of involving only a single real-valued parameter. The relationship between Bayesian and frequentist methods becomes more difficult, and more intriguing, when we venture into multi-dimensional situations.

Suppose then that instead of (8) the observed data $\mathbf{x}$ is a p-dimensional normal vector having unknown mean vector $\boldsymbol{\mu}$ and identity covariance matrix I,

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, I). \tag{19}$$

We wish to make inference about $\boldsymbol{\mu}$ having observed $\mathbf{x}$. Another way to express (19) is to say that we have p independent versions of (8), each with its own unknown parameter μ_i,

$$x_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \quad i = 1, 2, \dots, p. \tag{20}$$

At first glance this looks easy enough. The maximum likelihood estimate of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \mathbf{x}$, i.e. $\widehat{\mu}_i = x_i$ for $i = 1, 2, \dots, p$. The obvious 90% confidence region for $\boldsymbol{\mu}$ is a sphere centered at $\mathbf{x}$, say

$$C = \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \mathbf{x}\|^2 < c\},$$

where c is the upper 90% point of a chi-square distribution with p degrees of freedom. The obvious "objective prior" $g(\boldsymbol{\mu}) = 1$ for $\boldsymbol{\mu}$ in $\mathcal{R}^p$ also leads to point estimate $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ as *a posteriori* mean and to C as the *a posteriori* Bayes 90% region, so Bayesian and frequentist methods agree with each other.

It came as a great surprise to statisticians that there is something wrong with these "obvious" estimates and confidence regions. Charles Stein, working with his graduate student Willard James in the early 1960's, produced an estimator $\widetilde{\boldsymbol{\mu}}$ uniformly superior to the MLE $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ in terms of expected squared error,

$$\widetilde{\boldsymbol{\mu}} = \left[1 - \frac{p-2}{\|\mathbf{x}\|^2}\right] \cdot \mathbf{x}. \tag{21}$$

The James-Stein theorem states simply that in dimensions $p \geq 3$,

$$E\{\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} < E\{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} \tag{22}$$

for every possible choice of the parameter vector $\boldsymbol{\mu}$. Similarly, one can find 90% confidence regions uniformly smaller than C.

Statisticians have a lot invested in the estimator $\widehat{\boldsymbol{\mu}} = \mathbf{x}$, (it underlies analysis of variance and regression theory among other things) so Stein's result was initially viewed as paradoxical. A great deal of subsequent effort has gone into understanding Stein's paradox and tapping its potential for improved estimation in high dimensions.

The difference in squared error seen in (22) can be enormous in realistic applied problems. An example of Stein estimation (applied in a binomial setting rather than (14), but basically the same procedure) appears in Table 1. The batting averages of 18 major league baseball players in their first 45 at bats of the 1970 season are given in column one. They play the role of MLE estimates $\widehat{\mu}_i$ for the underlying true averages μ_i. The second column shows each player's average for the remainder of the season. These typically involved several hundred more at bats so we can use them as surrogates for the true μ_i. Column three gives a version of the James-Stein estimate $\widetilde{\boldsymbol{\mu}}$ (based only on the first 45 at bats) applicable to binomial data. The ratio it squared errors is seen to be overwhelming in the case, $\sum_{i=1}^{n}(\widehat{\mu}_i - \mu_i)^2 \Big/ \sum_{i=1}^{n}(\widetilde{\mu}_i - \mu_i)^2 = 3.50$.

Stein's work was carried out in a frequentist framework but results like (22) also disturbed Bayesians. Why can the estimates and confidence regions derived from the "uniformative" prior $g(\mu) \equiv 1$ be uniformly dominated? It turns out that there are better uninformative priors for high-dimensional problems.

Suppose we assume a scaled multivariate normal prior for $\boldsymbol{\mu}$ in (19),

$$\boldsymbol{\mu} \sim N_p(\mathbf{0}, A \cdot I), \tag{23}$$

but then take the scaler A in (23) to itself have a flat "prior prior", say

$$h(A) \equiv 1 \quad \text{for} \quad A > 0. \tag{24}$$

The Bayes estimator $\widehat{\mu}_{\text{Bayes}} = E\{\boldsymbol{\mu}|\mathbf{x}\}$ computed from the *hierarchical prior* (23)-(24) turns out to closely resemble the James-Stein estimator, and likewise dominates the MLE $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ as in (22). Another way to say this is that $\widehat{\boldsymbol{\mu}}_{\text{Bayes}}$ from (23)-(24) has smaller Bayes risk than $\widehat{\boldsymbol{\mu}}_{\text{Bayes}}$ from $g(\boldsymbol{\mu}) \equiv 1$ *no matter what the true prior may be.* In other words (23)-(24) is more uninformative than a flat prior, at least in dimensions $p \geq 3$. Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions.

	$\widehat{\mu}_i$	μ_i	$\widetilde{\mu}_i$
	First	Remainder	James-
	45	Season	Stein
Clemente	0.400	0.346	0.290
F. Robinson	0.378	0.298	0.286
F. Howard	0.356	0.276	0.282
Johnstone	0.333	0.222	0.277
Berry	0.311	0.273	0.273
Spencer	0.311	0.270	0.273
Kessinger	0.289	0.263	0.268
Alvarado	0.267	0.210	0.264
Santo	0.244	0.269	0.259
Swoboda	0.244	0.230	0.259
Unser	0.222	0.264	0.254
Williams	0.222	0.256	0.254
Scott	0.222	0.303	0.254
Petrocelli	0.222	0.264	0.254
Rodriguez	0.222	0.226	0.254
Campaneris	0.200	0.285	0.249
Munson	0.178	0.316	0.244
Alvis	0.156	0.200	0.239

Table 1: A baseball example of Stein estimation, 1970 season. The James-Stein estimator based on each player's first 45 at bats does move better than their observed averages at predicting subsequent performance.

Hierarchical priors dominate the recent Bayesian literature. They lead to an interesting amalgam of frequentist and Bayesian thinking called *empirical Bayes*. Empirical Bayes ideas neatly motivate the James-Stein estimation (21). For a given value of A in the normal prior (23), the *a posteriori* expectation of $\boldsymbol{\mu}$ given $\mathbf{x} \sim N_p(\boldsymbol{\mu}, I)$ is

$$\widehat{\boldsymbol{\mu}}_A = B\mathbf{x} \quad \text{where} \quad B = A/(A+1).$$

If we don't know A, we can replace B with the estimate

$$\widehat{B} = 1 - (p-2)/\|\mathbf{x}\|^2,$$

which is unbiased for B, in the usual frequentist sense, under assumptions (23) and (19).

In this way the James-Stein estimator $\widetilde{\boldsymbol{\mu}} = \widehat{B}\mathbf{x}$ is an obvious frequentist estimate for the unavailable Bayes rule $\widehat{\boldsymbol{\mu}}_A$ we would like to use.

Notice that the James-Stein estimator $\widetilde{\mu}_i$ for μ_i depends on the other observation $x_j,\ j \neq i$. Even though the component problems are assumed independent in (20), putting them together as in (21) is always better, in an expected total squared-error sense, than using the separate estimators $\widehat{\mu}_i$. Originally this seemed the most paradoxical element of James-Stein estimation. The empirical Bayes argument helps reduce the sense of paradox. Sets of problems that are analyzed together, like the baseball averages, may relate to each other at the parameter level even if their individual statistical errors are independent.

The twenty First Century has brought statisticians bigger problems to deal with, involving inferences for hundreds and even thousands of parameters at once. An uneasy alliance between Bayesian and frequentist methods, epitomized by empirical Bayes, seems to be replacing the time-worn adversarial relationship. We can hope all of this will all be settled by the time of phystat2103.

References

Mandelkern, M. (2002). “Setting Confidence Intervals for Bounded Parameters” *Statistical Science* **17** 149-192. [A statisticians' guide to the “neutrino problem” as it actually occurs.]

Feldman, G. and Cousins, R. (1998). “Unified approach to the classical statistical analysis of small signals” *Physical Review D* **57** 3873-89. [The most popular approach to bounded parameter confidence limits; considers lower as well as upper bounds.]

Efron, B. (1998). “R.A. Fisher in the 21st Century” *Statistical Science* **13** 95-122. [Discusses correctness and the frequentist-Bayesian relationship.]

Efron, B. and Hinkley, D. (1978). “Assessing the accuracy of the MLE: observed versus expected Fisher information” *Biomerika* **65** 457-87. [The Cauchy ancillarity example.]

Efron, B. and Gous, A. (2001). “Scales of evidence for model selection: Fisher versus Jeffreys” *IMS Lecture Series* **38** 208-256. [Jeffreys, a geophysicist and objective Bayesian, suggested a different scale of evidence than the familiar Fisherian .05 etc.]

Efron, B. and Morris, C. (1975). “Data analysis using Stein's estimator and its generalizations” *Jour. Amer. Stat. Assoc.* **70** 311-319. [Includes the baseball example. A popular version appeared in 1977 *Scientific American* **236** 119-127.]

Efron, B. “Robbins, Empirical Bayes, and Microarrays” *Annal. Stat.* **31** 366-378. [Empirical Bayes and high-dimensional problems in biogenetics.]