Statistical Science
2011, Vol. 26, No. 3, 299–316
DOI: 10.1214/11-STS352
© Institute of Mathematical Statistics, 2011

Is Bayes Posterior just Quick and Dirty Confidence?[1]

D. A. S. Fraser

Abstract. Bayes [*Philos. Trans. R. Soc. Lond.* **53** (1763) 370–418; **54** 296–325] introduced the observed likelihood function to statistical inference and provided a weight function to calibrate the parameter; he also introduced a confidence distribution on the parameter space but did not provide present justifications. Of course the names likelihood and confidence did not appear until much later: Fisher [*Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **222** (1922) 309–368] for likelihood and Neyman [*Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **237** (1937) 333–380] for confidence. Lindley [*J. Roy. Statist. Soc. Ser. B* **20** (1958) 102–107] showed that the Bayes and the confidence results were different when the model was not location. This paper examines the occurrence of true statements from the Bayes approach and from the confidence approach, and shows that the proportion of true statements in the Bayes case depends critically on the presence of linearity in the model; and with departure from this linearity the Bayes approach can be a poor approximation and be seriously misleading. Bayesian integration of weighted likelihood thus provides a first-order linear approximation to confidence, but without linearity can give substantially incorrect results.

Key words and phrases: Bayes, Bayes error rate, confidence, default prior, evaluating a prior, nonlinear parameter, posterior, prior.

1. INTRODUCTION

Statistical inference based on the observed likelihood function was initiated by Bayes (1763). This was, however, without the naming of the likelihood function or the apparent recognition that likelihood $L^0(\theta) = f(y^0; \theta)$ directly records the amount of probability at an observed data point y^0; such appeared much later (Fisher, 1922).

Bayes' proposal applies directly to a model with translation invariance that in current notation would be written $f(y - \theta)$; it recommended that a weight function or mathematical prior $\pi(\theta)$ be applied to the likelihood $L(\theta)$, and that the product $\pi(\theta)L(\theta)$ be treated as if it were a joint density for (θ, y). Then with observed data y^0 and the use of the conditional probability lemma, a posterior distribution $\pi(\theta|y) = c\pi(\theta)L^0(\theta)$ was obtained; this was viewed as a description of possible values for θ in the presence of data $y = y^0$. For the location model, as examined by the Bayes approach, translation invariance suggests a constant or flat prior $\pi(\theta) = c$ which leads to the posterior distribution $\pi(\theta|y^0) = f(y^0 - \theta)$ and, in the scalar case, gives the posterior survival probability $s(\theta) = \int_\theta^\infty f(y^0 - \alpha)\, d\alpha$, recording alleged probability to the right of a value θ.

The probability interpretation that would seemingly attach to this conditional calculation is as follows: if the θ values that might have been present in the application can be viewed as coming from the frequency pattern $\pi(\theta)$ with each θ value in turn giving rise to a y value in accord with the model and if the resulting y values that are close to y^0 are examined, then the associated θ values have the pattern $\pi(\theta|y^0)$.

D. A. S. Fraser is Professor Emeritus, Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3 and Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7 (e-mail: dfraser@utstat.toronto.edu).

[1] Discussed in 10.1214/11-STS352A, 10.1214/11-STS352B, 10.1214/11-STS352C and 10.1214/11-STS352D; rejoinder at 10.1214/11-STS352REJ.

The complication is that $\pi(\theta)$ as proposed is a mathematical construct and, correspondingly, $\pi(\theta|y^0)$ is just a mathematical construct. The argument using the conditional probability lemma does not produce probabilities from no probabilities: the probability lemma when invoked for an application has two distributions as input and one distribution as output; and it asserts the descriptive validity of the output on the basis of the descriptive validity of the two inputs; if one of the inputs is absent and an artifact is substituted, then the lemma says nothing, and produces no probabilities. Of course, other lemmas and other theory may offer something appropriate.

We will see, however, that something different is readily available and indeed available without the special translation invariance. We will also see that the procedure of augmenting likelihood $L^0(\theta)$ with a modulating factor that expresses model structure is a powerful first step in exploring information contained in Fisher's likelihood function.

An alternative to the Bayes proposal was introduced by Fisher (1930) as a confidence distribution. For the scalar-parameter case we can record the percentage position of the data point y^0 in the distribution having parameter value θ,

$$p(\theta) = p(\theta; y^0) = \int_{-\infty}^{y^0} f(y - \theta)\, dy.$$

This records the proportion of the θ population that is less than the value y^0. For a general data point y we have of course that $p(\theta; y)$ is uniformly distributed on $(0, 1)$, and, correspondingly, $p(\theta)$ from the data y^0 gives the upper-tail distribution function or survivor function for confidence, as introduced by Fisher (1935). A basic way of presenting confidence is in terms of quantiles. If we set $p(\theta) = 0.95$ and solve for θ, we obtain $\theta = \hat{\theta}_{0.95}$ which is the value with right tail confidence 95% and left tail confidence 5%; this would typically be called the 95% lower confidence bound, and $(\hat{\theta}_{0.95}, \infty)$ would be the corresponding 95% confidence interval.

For two-sided confidence the situation has some subtleties that are often overlooked. With the large data sets that have come from the colliders of High Energy Physics, a Poisson count can have a mean at a background count level or at a larger value if some proposed particle is actually present. A common practice in the High Energy Physics literature (Mandelkern, 2002) has been to form two-sided confidence intervals and to allow the confidence contributions in the two tails to be different, thereby accommodating some optimality criterion; see also some discussion in Section 4. In practice, this meant that the confidence lower bound shied away from the critical parameter lower bound describing just the background radiation. This mismanaged the detection of a new particle. Accordingly, our view is that two-sided intervals should typically have equal or certainly designated amounts of confidence in the two tails. With this in mind, we now restrict the discussion to the analysis of the confidence bounds as described in the preceding paragraph and view confidence intervals as being properly built on individual confidence bounds with designated confidence values.

As a simple example consider the Normal(μ, σ_0^2), and let $\phi(z)$ and $\Phi(z)$ be the standard normal density and distribution functions. The p-value from data y^0 is

$$p(\mu) = \int_{-\infty}^{y^0} \phi\left(\frac{y-\mu}{\sigma_0}\right) dy = \Phi\left(\frac{y^0-\mu}{\sigma_0}\right),$$

which has normal distribution function shape dropping from 1 at $-\infty$ to 0 at $+\infty$; it records the probability position of the data with respect to a possible parameter value μ; see Figure 1(a). From the confidence viewpoint, $p(\mu)$ is recording the right tail confidence distribution function, and the confidence distribution is Normal(y^0, σ_0^2).

The Bayes posterior distribution for μ using the invariant prior has density $c\phi\{(y^0 - \mu)/\sigma_0\}$; this is Normal$(y^0, \sigma_0^2)$. The resulting posterior survivor function value is

$$s(\mu) = \int_{\mu}^{\infty} \phi\left(\frac{y^0-\alpha}{\sigma_0}\right) d\alpha = \Phi\left(\frac{y^0-\mu}{\sigma_0}\right)$$

and its values are indicated in Figure 1(b); the function provides a probability-type evaluation of the right tail interval (μ, ∞) for the parameter. For this we have

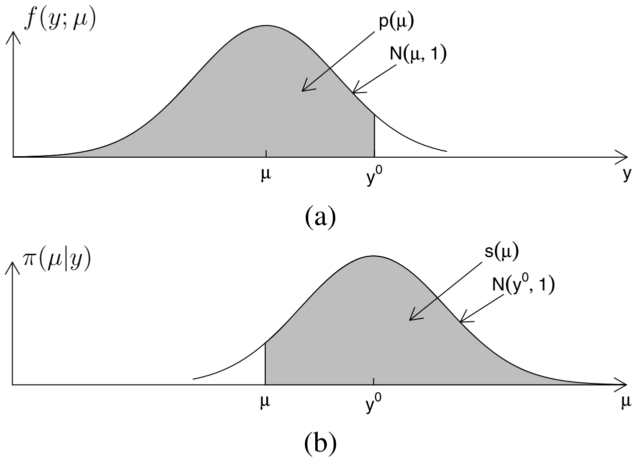


FIG. 1. Normal$(\mu, 1)$ *model*: *The density of* y *given* μ *in* (a); *the posterior density of* μ *given* y^0 *in* (b). *The* p*-value* $p(\mu)$ *from* (a) *is equal to the survivor value* $s(\mu)$ *in* (b).

used the letter s to suggest the “survivor” aspect of the Bayes analogue of the present one-sided frequentist p-value.

For a second example consider the model $y = \theta + z$, where z has the standard extreme value distribution with density $g(z) = e^{-z}\exp\{-e^{-z}\}$ and distribution function $G(z) = \exp(-e^{-z})$. The p-value from data y^0 is

$$p(\theta) = \int_{-\infty}^{y^0} g(y-\theta)\,dy = G(y^0 - \theta)$$
$$= \exp\{-e^{-(y^0-\theta)}\},$$

which records the probability position of the data in the θ distribution; it can be viewed as a right tail distribution function for confidence. The posterior distribution for θ using the Bayes invariant prior has density $g(y^0 - \theta)$ and can be described as a reversed extreme value distribution centered at y^0. The posterior survivor function is

$$s(\theta) = \int_{\theta}^{\infty} g(y^0 - \alpha)\,d\alpha = \exp\{-e^{-(y^0-\theta)}\},$$

and again agrees with the p-value $p(\theta)$; see Figure 2.

Of course, in general for a location model $f(y-\theta)$ as examined by Bayes, we have

$$p(\theta) = \int_{-\infty}^{y^0} f(y-\theta)\,dy = \int_{-\infty}^{y^0-\theta} f(z)\,dz$$
$$= \int_{\theta}^{\infty} f(y^0-\alpha)\,d\alpha = s(\theta),$$

and, thus, the Bayes posterior distribution is equal to the confidence distribution. Or, more directly, the Bayes posterior distribution is just standard confidence.

Lindley (1958) presented this result and under suitable change of variable and parameter showed more: that the p-value and s-value are equal if and only if the model $f(y;\theta)$ is a location model $f(y-\theta)$. In his perspective then, this argued that the confidence approach was flawed, confidence as obtained by inverting the p-value function as a pivot. From a different perspective, however, it argues equally that the Bayes approach is flawed, and does not have the support of the confidence interpretation unless the model is location.

Lindley objected also to the term probability being attached to the original Fisher word for confidence, viewing probability as appropriate only in reference to the conditional type calculations used by Bayes. By contrast, repetition properties for confidence had been clarified by Neyman (1937). As a consequence, in the discipline of statistics, the terms probability and distribution were then typically not used in the confidence context, but were in the Bayes context. The repetition properties, however, do not extend to the Bayes calculation except for simple location cases, as we will see; but they do extend for the confidence inversion. We take this as strong argument that the term probability is less appropriate in the Bayesian weighted likelihood context than in the confidence inversion context.

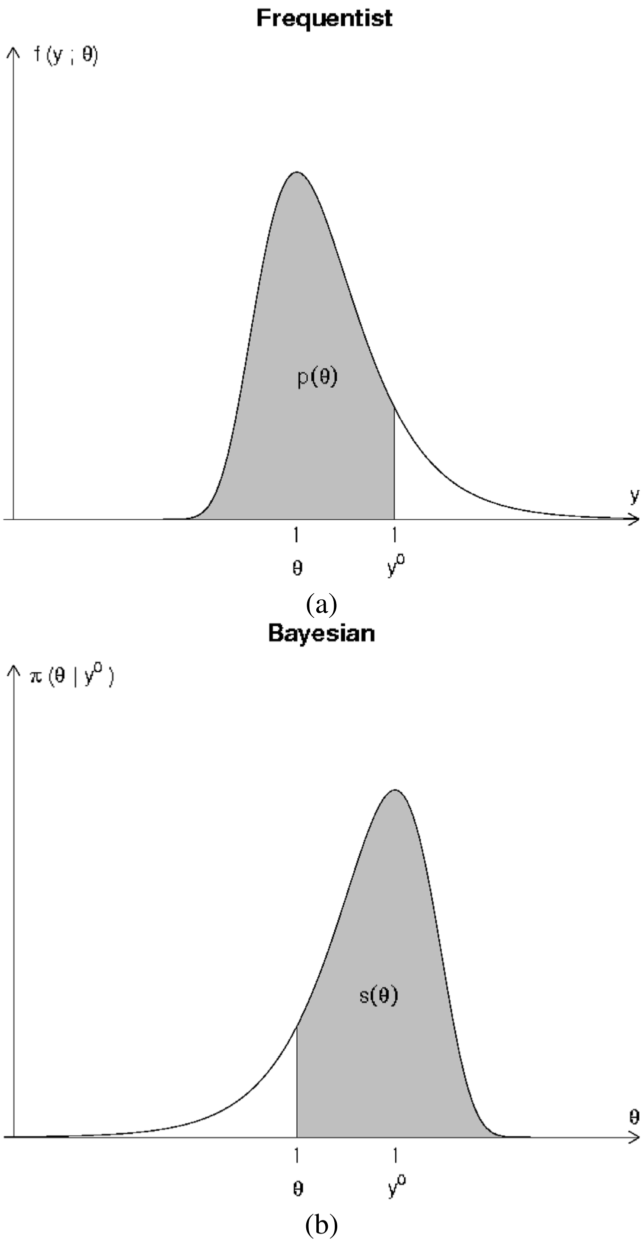


FIG. 2. *The extreme value* EV$(\theta, 1)$ *model: the density of y given θ in* (a); *the posterior density of θ given y^0 in* (b). *The p-value $p(\theta)$ from* (a) *is equal to the survivor value $s(\theta)$ in* (b).

The location model, however, is extremely special in that the parameter has a fundamental linearity and this linearity is expressed in the use of the flat prior with respect to the location parameter. Many extensions of the Bayes mathematical prior have been proposed trying to achieve the favorable behavior of the original Bayes, for example, Jeffreys (1939, 1946) and Bernardo (1979). We refer to such priors as default priors, priors to elicit information from an observed likelihood function. And we will show that if the parameter departs from a basic linearity, then the Bayes posterior can be seriously misleading. Specifically, we will show that with moderate departures from linearity the Bayes calculation can give an acceptable approximation to confidence, but that with more extreme departure from linearity or with large parameter dimension it can give unacceptable approximations.

John Tukey actively promoted a wealth of simple statistical methods as a means to explore data; he referred to them as quick and dirty methods. They were certainly quick using medians and ranges and other easily accessible characteristics of data. And they were dirty in the sense of ignoring characteristics that in the then currently correct view were considered important. We argue that Bayes posterior calculations can appropriately be called quick and dirty, quick and dirty confidence.

There are also extensions of the Bayes approach allowing the prior to reflect the viewpoint or judgment or prejudice of an investigator; or to reflect the elicited considerations of those familiar with the context being investigated. Arguments have been given that such a viewpoint or consideration can be expressed as probability; but the examples that we present suggest otherwise.

There are of course contexts where the true value of the parameter has come from a source with a known distribution; in such cases the prior is real, it is objective, and could reasonably be considered to be a part of an enlarged model. Then whether to include the prior becomes a modeling issue. Also, in particular contexts, there may be legal, ethical or moral issues as to whether such outside information can be included. If included, the enlarged model is a probability model and accordingly is not statistical: as such, it has no statistical parameters in the technical sense and thus predates Bayes and can be viewed as being probability itself not Bayesian. Why this would commonly be included in the Bayesian domain is not clear; it is not indicated in the original Bayes, although it was an area neglected by the frequentist approach. Such a prior describing a known source is clearly objective and can properly be called an objective prior; this conflicts, however, with some recent Bayesian usage where the term objective is misapplied and refers to the mathematical priors that we are calling default priors.

In Section 2 we consider the scalar variable scalar parameter case and determine the default prior that gives posteriors with reliable quantiles; some details for the vector parameter case are also discussed. In Section 3 we argue that the only satisfactory way to assess distributions for unobserved quantities is by means of the quantiles of such distributions; this provides the basis then for comparing the Bayesian and frequentist approaches.

In Sections 4–6 we explore a succession of examples that examine how curvature in the model or in the parameter of interest can destroy any confidence reliability in the default Bayes approach, and thus in the Bayesian use of just likelihood to present a distribution purporting to describe an unknown parameter value.

In Sections 7 and 8 we discuss the merits of the conditional probability formula when used with a mathematical prior and also the merits of the optimality approach; then Section 9 records a brief discussion and Section 10 a summary.

2. BUT IF THE MODEL IS NONLINEAR

With a location model the confidence approach gives $p(\theta)$ and the default Bayes approach gives $s(\theta)$, and these are equal. Now consider things more generally and initially examine just a statistical model $f(y;\theta)$ where both y and θ are scalar or real valued as opposed to vector valued, but without the assumed linear relationship just discussed.

Confidence is obtained from the observed distribution function $F^0(\theta)$ and a posterior is obtained from the observed density function $f^0(\theta)$. For convenience we assume minimum continuity and that $F(y;\theta)$ is stochastically increasing and attains both 0 and 1 under variation in y or θ. The confidence p-value is directly the observed distribution function,

$$p(\theta) = F^0(\theta) = F(y^0;\theta),$$

which can be rewritten mechanically as

$$p(\theta) = \int_\theta^\infty -F_{;\theta}(y^0;\alpha)\,d\alpha;$$

the subscript denotes partial differentiation with respect to the corresponding argument. The default Bayes s-value is obtained from likelihood, which is the observed density function $f(y^0;\theta) = F_y(y^0;\theta)$:

$$s(\theta) = \int_\theta^\infty \pi(\alpha)F_y(y^0;\alpha)\,d\alpha.$$

If $p(\theta)$ and $s(\theta)$ are in agreement, then the direct comparison of the integrals implies that

$$\pi(\theta) = \frac{-F_{;\theta}(y^0;\theta)}{F_y(y^0;\theta)}.$$

This presents $\pi(\theta)$ as a possibly data-dependent prior. Of course, data dependent priors have a long but rather infrequent presence, for example, Box and Cox (1964), Wasserman (2000) and Fraser et al. (2010b). The preceding expression for the prior can be rewritten as

$$\pi(\theta) = \left.\frac{dy}{d\theta}\right|_{y^0}$$

by directly differentiating the quantile function $y = y(u,\theta)$ for fixed p-value to $u = F(y;\theta)$ and taking the observed value, or by taking the total differential of $F(y;\theta)$; Lindley's (1958) result follows by noting that the differential equation $dy/d\theta = a(\theta)/b(y)$ integrates to give a location model.

Now suppose we go beyond the simple case of the scalar model and allow that y is a vector of length n and θ is a vector of length p. In many applications $n > p$; but here we assume that dim y has been reduced to p by conditioning (see, e.g., Fraser, Fraser and Staicu, 2010c), and that a smooth pivot $z(y,\theta)$ with density $g(z)$ describes how the parameter affects the distribution of the variable. The density for y is available by inverting from pivot to sample space:

$$g(z)dz = f(y;\theta)\,dy = g\{z(y;\theta)\}|z_y(y;\theta)|\,dy,$$

where the subscript again denotes partial differentiation.

For confidence a differential element is obtained by inverting from pivot to parameter space:

$$g(z)dz = g\{z(y^0;\theta)\}|z_{;\theta}(y^0;\theta)|\,d\theta.$$

And for posterior probability the differential element is obtained as weighted likelihood

$$g(z)dz = g\{z(y^0;\theta)\}|z_y(y^0;\theta)|\pi(\theta)\,d\theta.$$

The confidence and posterior differential elements are equal if

$$\pi(\theta) = \frac{|z_{;\theta}(y^0;\theta)|}{|z_y(y^0;\theta)|};$$

we call this the default prior for the model $f(y;\theta)$ with data y^0. As $dy/d\theta = z_y^{-1}(y^0;\theta)z_{;\theta}(y^0;\theta)$ for fixed z, we will have confidence equal to posterior if $\pi(\theta) = |dy/d\theta|_{y^0}|$, a simple extension of the scalar case. The matrix $dy/d\theta|_{y^0}$ can be called the sensitivity of the parameter at the data point y^0 and the determinant provides a natural weighting or scaling function $\pi(\theta)$ for the parameter; this sensitivity is just presenting how parameter change affects the model and is recording this just at the relevant point, the observed data.

3. HOW TO EVALUATE A POSTERIOR DISTRIBUTION

(i) *Distribution function or quantile function*. In the scalar parameter case, both $p(\theta)$ and $s(\theta)$ have the form of a right tail distribution function or survivor function. In the Bayesian framework, the function $s(\theta)$ is viewed as a distribution of posterior probability. In the frequentist framework, the function $p(\theta)$ can be viewed as a distribution of confidence, as introduced by Fisher (1930) but originally called fiducial; it has long been a familiar theme, frequentist or Bayesian, that it is inappropriate to treat such a function as a distribution describing possible values for the true parameter.

For a scalar parameter model with data, the Bayes and the confidence approaches with data each lead to a probability-type evaluation on the parameter space; and these can be different as Lindley (1958) demonstrated and as we have quantified in the preceding section. Surely then, they both cannot be correct. So, how to evaluate such posterior distributions for the parameter?

A probability description is a rather complex thing even for a scalar parameter: ascribing a probability-type assessment to one-sided intervals, two-sided intervals, and more general sets. What seems more tangible but, indeed, is equivalent is to focus on the reverse, the quantiles: choose an amount β of probability and then determine the corresponding quantile $\hat{\theta}_\beta$, a value with the alleged probability $1-\beta$ to the left and with β to the right. We then have that a particular interval $(\hat{\theta}_\beta,\infty)$ from the data has the alleged amount β. Here we focus on such quantiles $\hat{\theta}_\beta$ on the scale for θ. In particular, we might examine the 95% quantile $\hat{\theta}_{0.95}$, the median quantile $\hat{\theta}_{0.50}$, the 5% quantile $\hat{\theta}_{0.05}$, and others, all as part of examining an alleged distribution for θ obtained from the data.

For the Normal(μ,σ_0^2) example with data y^0, the confidence approach gives the β-level quantile

$$\hat{\mu}_\beta = y^0 - z_\beta\sigma_0,$$

where $\Phi(z_\beta) = \beta$ as based on the standard normal distribution function Φ. In particular, the 95%, 50% and

5% quantiles are

$$\hat{\mu}_{0.95} = y^0 - 1.64\sigma_0, \quad \hat{\mu}_{0.50} = y^0,$$
$$\hat{\mu}_{0.05} = y^0 + 1.64\sigma_0;$$

and the corresponding confidence intervals are

$$(y^0 - 1.64\sigma_0, \infty), \quad (y^0, \infty), \quad (y^0 + 1.64\sigma_0, \infty),$$

with the lower confidence bound in each case recording the corresponding quantile.

Now more generally suppose we have a model $f(y;\theta)$ and data y^0, and that we want to evaluate a proposed procedure, Bayes, frequentist or other, that gives a probability-type evaluation of where the true parameter θ might be. As just discussed, we can focus on some level, say, β, and then examine the corresponding quantile $\hat{\theta}_\beta$ or the related interval $(\hat{\theta}_\beta, \infty)$. In any particular instance, either the true θ is in the interval $(\hat{\theta}_\beta, \infty)$, or it is not. And yet the procedure has put forward a numerical level β for the presence of θ in $(\hat{\theta}_\beta, \infty)$. What does the asserted level β mean?

(ii) *Evaluating a proposed quantile*. The definitive evaluation procedure is in the literature: use a Neyman (1937) diagram. The model $f(y;\theta)$ sits on the space $S \times \Omega$ which here is the real line for S *cross* the real line for Ω; this is just the plane R^2. For any particular y the procedure gives a parameter interval $(\hat{\theta}_\beta(y), \infty)$; if we then assemble the resulting intervals, we obtain a region

$$A_\beta = \bigcup \{y\} \times (\hat{\theta}(y), \infty) = \{(y,\theta) : \theta \text{ in } (\hat{\theta}(y), \infty)\}$$

on the plane. For the confidence procedure in the simple Normal$(\theta, 1)$ case, Figure 3 illustrates the 97.5% quantile $\hat{\theta}_{0.975}$ for that confidence procedure; the region $A_\beta = A_{0.975}$ is to the upper left of the angled line and it represents the $\beta = 97.5\%$ allegation concerning the true θ, as proceeding from the confidence procedure.

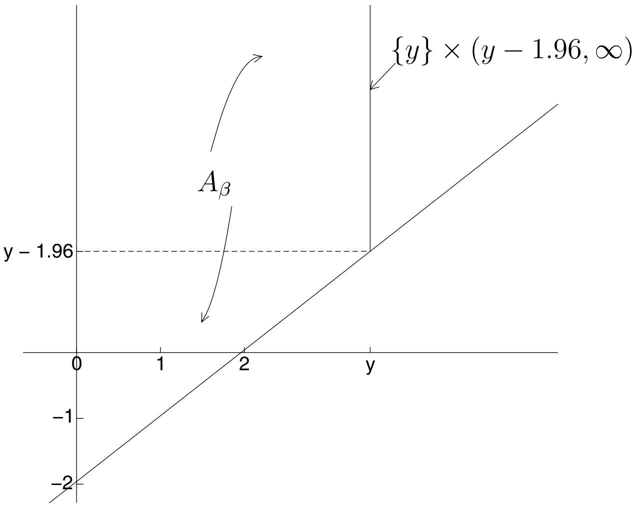


FIG. 3. *The* 97.5% *allegation for the* Normal *confidence procedure, on the* (y,θ)*-space.*

Now, more generally for a scalar parameter, we suggest that the sets A_β present precisely the essence of a posterior procedure: how the procedure presents information concerning the unknown θ value. We can certainly examine these for various values of β and thus investigate the merits of any claim implicit in the alleged levels β.

The level β is attached to the claim that θ is in $(\hat{\theta}_\beta(y), \infty)$, or, equivalently, that (y, θ) is in the set A_β. In any particular instance, there is of course a true value θ, and either it is in $\{\hat{\theta}_\beta(y), \infty\}$ or it is not in $\{\hat{\theta}_\beta(y), \infty\}$. And the true θ did precede the generation of the observed y in full accord with the probabilities given by the model. Accordingly, a value θ for the parameter in the model implies an actual Proportion of true assertions consequent to that θ value:

$$\text{Propn}(A_\beta; \theta) = \Pr\{A_\beta \text{ includes } (y,\theta); \theta\}.$$

This allows us to check what relationship the actual Proportion bears to the value β asserted by the procedure: is it really β or is it something else?

Of course, there may be contexts where in addition we have that the θ value has been generated by some random source described by an available prior density $\pi(\theta)$, and we would be interested in the associated Proportion,

$$\text{Propn}(A_\beta; \pi) = \int \text{Propn}(A_\beta; \theta)\pi(\theta)\, d\theta,$$

presenting the average relative to the source density $\pi(\theta)$.

(iii) *Comparing proposed quantiles*. For the Bayes procedure with the special original linear model $f(y - \theta)$ we have by the usual calculations that

$$\text{Propn}(A_\beta; \theta) \equiv \beta$$

for all θ and β: the alleged level β agrees with the actual Proportion of true assertions that are made. And, more generally, if the θ value has been generated by a source $\pi(\theta)$, then it follows that the alleged level β does agree with the actual Proportion: thus, $\text{Propn}(A_\beta; \pi) \equiv \beta$.

For the standard confidence procedure in the context of an arbitrary continuous scalar model $f(y;\theta)$, we have by the standard calculations that

$$\text{Propn}(A_\beta; \theta) \equiv \Pr\{(y,\theta) \text{ in } A_\beta; \theta\}$$
$$\equiv \Pr\{F(y;\theta) \le \beta; \theta\} \equiv \beta$$

for all θ and β. Of course, in the special Bayes location model $f(y-\theta)$ the Bayes original procedure does coincide with the confidence procedure: the original Bayes was confidence in disguise.

Now for some proposed procedure having a region A_β with alleged level β, there is of course the possibility that the actual Proportion is less than β for some θ and is greater than β for some other θ and yet when averaged with a particular prior $\pi(\theta)$ gives a revised Propn$(A_\beta;\pi)$ that does have the value β; the importance or otherwise of this we will discuss later.

But we now ask, what is the actual Proportion for a Bayes procedure in nonlocation models? Toward this, we next examine a succession of examples where the linearity is absent to varying degrees, where the parameter to variable relationship is nonlinear!

4. NONLINEARITY AND BOUNDED PARAMETER: THE ERRORS ARE $O(1)$

We first examine an extreme form of nonlinearity, where the range of the parameter is bounded. This is a familiar problem in the current High Energy Physics of particle accelerators and the related search and detection of possible new particles: a particle count has a Poisson(θ) distribution but θ is bounded below by θ_0, which represents the contribution from background radiation. For some discussion see Mandelkern (2002), Reid and Fraser (2003) and Fraser, Reid and Wong (2004).

The critical issues are more easily examined in a continuous context. For this, suppose that y is Normal(θ,σ_0^2) and that it is known that $\theta \ge \theta_0$ with an interest in detecting whether θ is actually larger than θ_0; let y^0 be the observed data point; this continuous version was also mentioned in Mandelkern (2002), Woodroofe and Wang (2000) and Zhang and Woodroofe (2003). For convenience here and without loss of generality, we take the known $\sigma_0 = 1$ and the lower bound $\theta_0 = 0$.

From a frequentist viewpoint, there is the likelihood

$$L^0(\theta) = c\phi(y^0-\theta)$$

recording probability at the data, again using $\phi(z)$ for the standard normal density. And also there is the p-value

$$p(\theta) = \Phi(y^0-\theta)$$

recording probability left of the data. They each offer a basic presentation of information concerning the parameter value θ; see Figure 4(a) and (b). Also note that

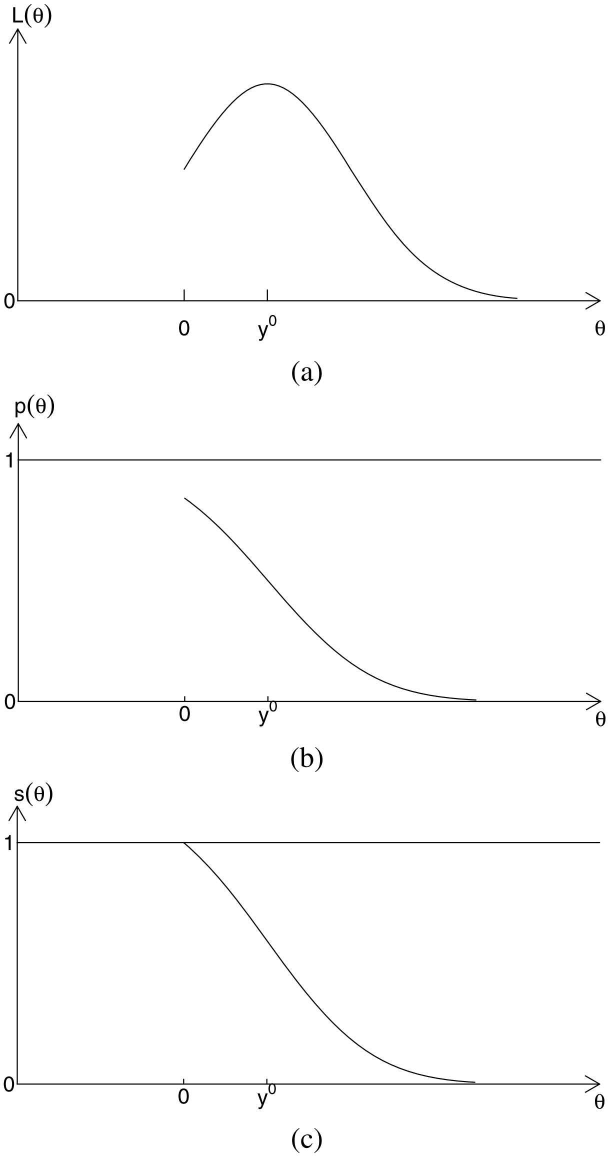


FIG. 4. *The* Normal$(\theta,1)$ *with* $\theta \ge \theta_0 = 0$: (a) *the likelihood function* $L(\theta)$; (b) *p-value function* $p(\theta) = \Phi(y^0-\theta)$; (c) *s-value function* $s(\theta) = \Phi(y^0-\theta)/\Phi(y^0)$.

$p(\theta)$ does not reach the value 1 at the lower limit θ_0 for θ; of course, the p-value is just recording the statistical position of the data y^0 under possible θ values, so there is no reason to want or expect such a limit.

First consider the confidence approach. The interval $(0,\beta)$ for the p-value function gives the interval $\{\max(\theta_0, y^0 - z_\beta), \infty\}$ for θ when we acknowledge the lower bound, or gives the interval $(y^0 - z_\beta, \infty)$ when we ignore the lower bound. In either case the actual Proportion is equal to the alleged value β, regardless of the true value of θ. There might perhaps be mild discomfort that if we ignore the lower bound and calculate the interval, then it can include parameter values that are not part of the problem; but nonetheless the alleged level is valid.

Now consider the default Bayes approach. The model $f(y;\theta) = \phi(y^0 - \theta)$ is translation invariant for $\theta \ge \theta_0$, and this would indicate the constant prior $\pi(\theta) = c$, at least for $\theta \ge \theta_0$. Combining the prior and likelihood and norming as usual gives the posterior density

$$\pi(\theta|y^0) = \frac{\phi(y^0 - \theta)}{\Phi(y^0)}, \quad \theta \ge 0,$$

and then gives the posterior survivor value

$$s(\theta) = \frac{\Phi(y^0 - \theta)}{\Phi(y^0)}, \quad \theta \ge 0.$$

See Figure 4(c). The β-quantile of this truncated normal distribution for θ is obtained by setting $s(\theta) = \beta$ and solving for θ:

$$\hat{\theta}_\beta = y^0 - z_{\beta\Phi(y^0)},$$

where again z_γ designates the standard normal γ-quantile.

We are now in a position to calculate the actual Proportion, namely, the proportion of cases where it is true that θ is in the quantile interval, or, equivalently, the proportion of cases where $(\hat{\theta}_\beta, \infty)$ includes the true θ value:

$$\begin{aligned}\mathrm{Propn}(\theta) &= \Pr\{y - z_{\beta\Phi(y)} < \theta : \theta\} \\ &= \Pr\{z < z_{\beta\Phi(\theta+z)}\},\end{aligned}$$

where z is taken as being Normal(0, 1); this expression can be written as in integral $\int_S \phi(z)\,dz$ with $S = \{z : \Phi(z) < \beta\Phi(\theta + z)\}$ and can routinely be evaluated numerically for particular values of θ and β. In particular, for θ at the lower limit $\theta = \theta_0 = 0$ the coverage set S becomes $S = \{z : \Phi(z) < \beta\Phi(z)\}$, which is clearly the empty set unless $\beta = 1$. In particular, at the lower limit $\theta = \theta_0 = 0$ the Propn(θ_0) has the phenomenal value zero, Propn$(\theta_0) = 0$, which is a consequence of the empty set just mentioned; certainly an unusual performance property for a claimed Bayes coverage of, say, β when, as typical, β is not zero.

In Figure 5 we plot this Proportion against β for the $\beta = 50\%$ quantile; and we note that it is uniformly less than the nominal, the claimed 50%. In particular, at the lower limit $\theta = \theta_0 = 0$ the Propn$(\theta_0) = 0$ has the phenomenal value zero, as mentioned in the preceding paragraph; certainly an unusual performance property for a claimed Bayes coverage of $\beta = 50\%$! Then in Figure 6 we plot the proportion for $\beta = 90\%$ and for $\beta = 10\%$; again we note that the actual Proportion is uniformly less than the claimed value, and again Propn(θ) has the extraordinary coverage value 0 when the parameter is at the lower bound 0. Of course, the departure would be in the other direction in the case of an upper bound.

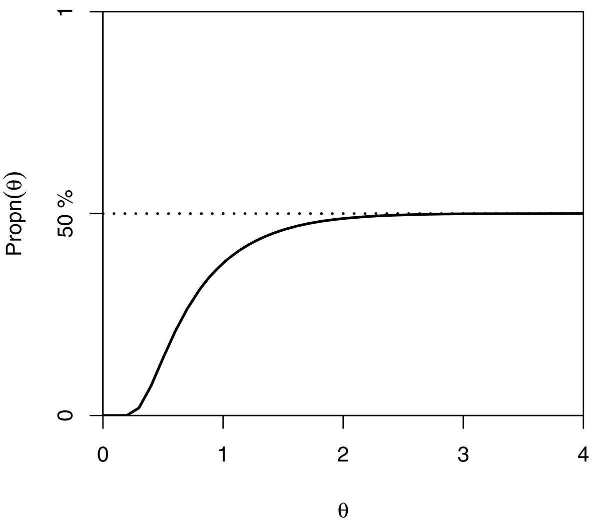


FIG. 5. Normal *with bounded mean*: *the actual Proportion for the claimed level $\beta = 50\%$ is strictly less than the claimed* 50%.

In summary, in a context with a bound on the parameter, the performance error with the Bayes calculation can be of asymptotic order $O(1)$.

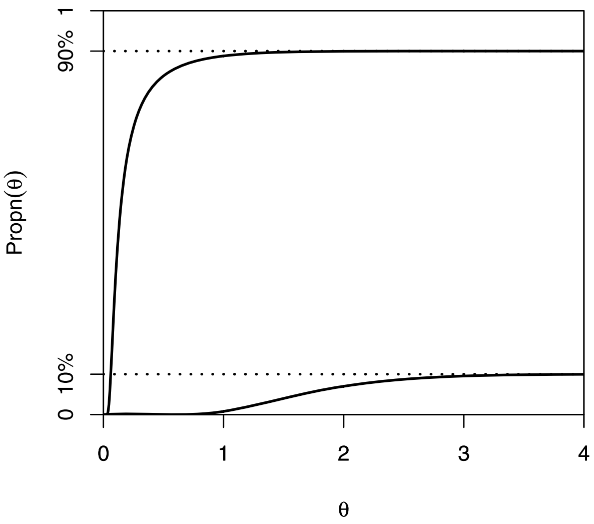


FIG. 6. Normal *with bounded mean*: *the actual Proportions for the claimed level $\beta = 90\%$ and $\beta = 10\%$ are strictly less than the claimed.*

5. NONLINEARITY AND PARAMETER CURVATURE: THE ERRORS ARE $O(n^{-1/2})$

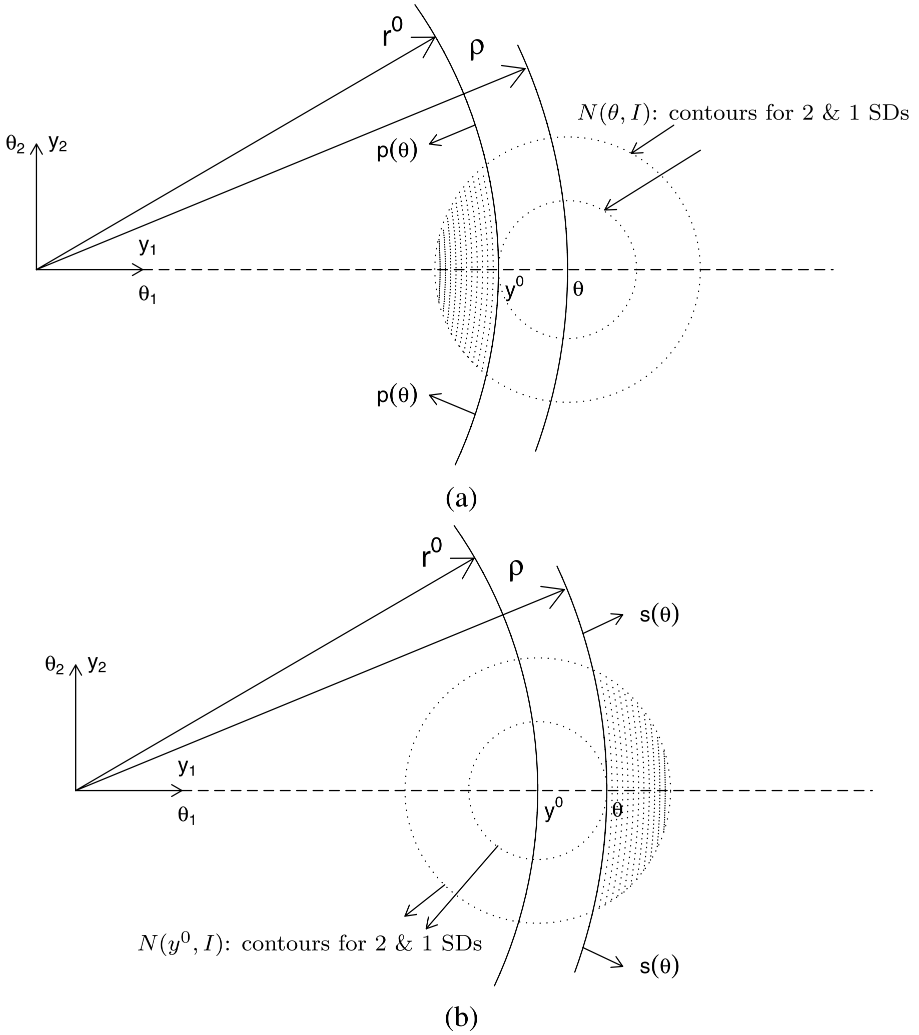
A bound on a parameter as just discussed is a rather extreme form of nonlinearity. Now consider a very direct and common form of curvature. Let (y_1, y_2) be Normal$(\theta; I)$ on R^2 and consider the quadratic interest parameter $(\theta_1^2 + \theta_2^2)$, or the equivalent $\rho(\theta) = (\theta_1^2 + \theta_2^2)^{1/2}$ which has the same dimensional units as the θ_i; and let $y^0 = (y_1^0, y_2^0)$ be the observed data. For asymptotic analysis we would view the present variables as being derived from some antecedent sample of size n and they would then have the Normal$(\theta, I/n)$ distribution.

From the frequentist view there is an observable variable $r = (y_1^2 + y_2^2)^{1/2}$ that in some pure physical sense measures the parameter ρ. It has a noncentral chi distribution with noncentrality ρ and degrees of freedom 2. For convenience we let $\chi_2(\rho)$ designate such a variable with distribution function $H_2(\chi, \rho)$, which is typically available in computer packages; and its square can be expressed as $\chi_2^2 = (z_1 + \rho)^2 + z_2^2$ in terms of standard normal variables and it has the noncentral chi-square distribution with 2 degrees of freedom and noncentrality usually described by ρ^2. The distribution of r is free of the nuisance parameter which can conveniently be taken as the polar angle $\alpha = \arctan(\theta_2/\theta_1)$. The resulting p-value function for ρ is

$$(1) \qquad p(\rho) = \Pr\{\chi_2(\rho) \le r^0\} = H_2(r^0; \rho).$$

See Figure 7(a), where for illustration we examined the behavior for $\theta = y^0 + 1$.

From the frequentist view there is the directly measured p-value $p(\rho)$ with a Uniform(0, 1) distribution, and any β level lower confidence quantile is available



(a)

(b)

FIG. 7. (a) *The model is $N(\theta; I)$; region for $p(\theta)$ is shown.* (b) *The posterior distribution for θ is $N(y^0; I)$; region for $s(\theta)$ is shown.*

immediately by solving $\beta = H_2(r^0; \rho)$ for ρ in terms of r^0.

From the Bayes view there is a uniform prior $\pi(\theta) = c$ as directly indicated by Bayes (1763) for a location model on the plane R^2. The corresponding posterior distribution for θ is then Normal$(y^0; I)$ on the plane. And the resulting marginal posterior for ρ is described by the generic variable $\chi_2(r^0)$. As r is stochastically increasing in ρ, we have that the Bayes analog of the p-value is the posterior survivor value obtained by an upper tail integration

$$(2) \qquad s(\rho) = \Pr\{\chi_2(r^0) \geq \rho\} = 1 - H_2(\rho; r^0).$$

The Bayes $s(\rho)$ and the frequentist $p(\rho)$ are actually quite different, a direct consequence of the obvious curvature in the parameter $\rho = (\theta_1^2 + \theta_2^2)^{1/2}$. The presence of the difference is easily assessed visually in Figure 7 by noting that in either case there is a rotationally symmetric normal distribution with unit standard deviation which is at the distance $d = 1$ from the curved boundary used for the probability calculations, but the curved boundary is cupped away from the Normal distribution in the frequentist case and is cupped toward the Normal distribution in the Bayes case; this difference is the direct source of the Bayes error.

From (1) and (2) we can evaluate the posterior error $s(\rho) - p(\rho) = 1 - H_2(\rho; r^0) - H_2(r^0; \rho)$ which is plotted against ρ in Figure 8 for $r^0 = 5$. This Bayes error here is always greater than zero. This happens widely with a parameter that has curvature, with the error in one or other direction depending on the curvature being positive or negative relative to increasing values of the parameter. Some aspects of this discrepancy are discussed in David, Stone and Zidek (1973) as a marginalization paradox.

Now in more detail for this example, consider the β lower quantile $\hat{\rho}_\beta$ of the Bayes posterior distribution for the interest parameter ρ. This β quantile for the parameter ρ is obtained from the $\chi_2(r^0)$ posterior distribution for ρ giving

$$\hat{\rho}_\beta = \chi_{1-\beta}(r^0),$$

where we now use $\chi_\gamma(r)$ for the γ quantile of the noncentral chi variable with 2 degrees of freedom and noncentrality r, that is, $H_2(\chi_\gamma; r) = \gamma$. We are now in a position to evaluate the Bayes posterior proposal for ρ. For this let Propn$(A_\beta; \theta)$ be the proportion of true assertions that ρ is in $A_\beta = \{\hat{\rho}_\beta(r), \infty\}$; we have

$$\begin{aligned}\text{Propn}(A_\beta; \rho) &= \Pr\{\rho \text{ in } (\hat{\rho}_\beta(r), \infty); \rho\}\\ &= \Pr\{\hat{\rho}_\beta(r) \leq \rho; \rho\}\\ &= \Pr\{\chi_{1-\beta}(r) \leq \rho; \rho\},\end{aligned}$$

where the quantile $\hat{\rho}_\beta(r)$ is seen to be the $(1-\beta)$ point of a noncentral chi variable with degrees of freedom 2 and noncentrality r, and the noncentrality r has a noncentral chi distribution with noncentrality ρ. The actual Proportion under a parameter value ρ can thus be presented as

$$\begin{aligned}\text{Propn}(A_\beta; \rho) &= \Pr[\chi_{1-\beta}\{\chi_2(\rho)\} \leq \rho; \rho]\\ &= \Pr[1 - \beta < H_2\{\rho; \chi_2(\rho)\}],\end{aligned}$$

which is available by numerical integration on the real line for any chosen β value.

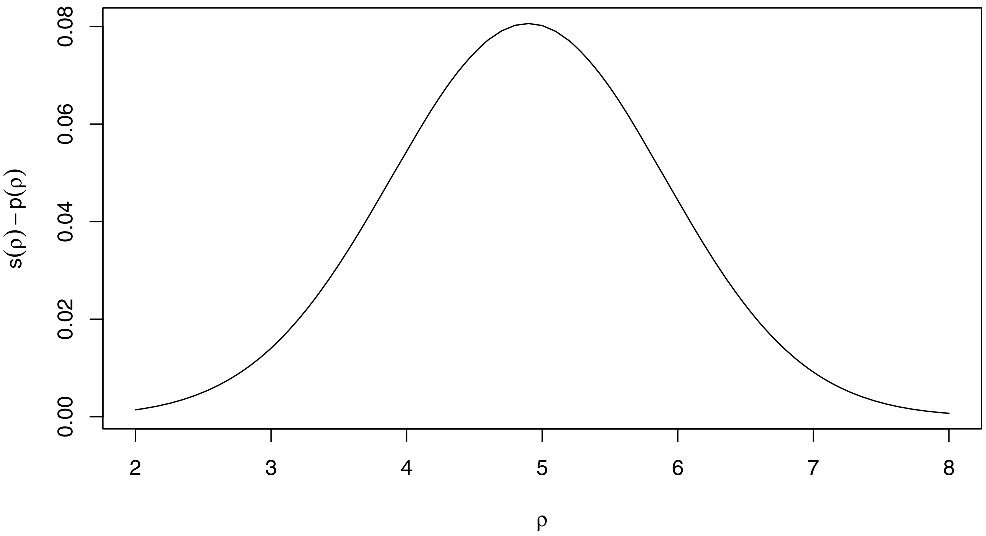


FIG. 8. *The Bayes error $s(\rho) - p(\rho)$ from the $N(\theta, I)$ model with data $y^0 = (5, 0)$.*

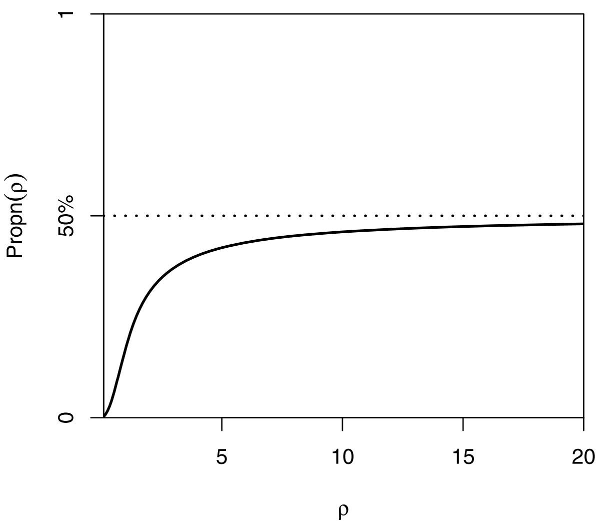


FIG. 9. *Proportion with claimed level $\beta = 50\%$.*

We plot the actual Propn$(A_{50\%}; \rho)$ against ρ in Figure 9 and note that it is always less than the alleged 50%. We then plot the Proportion for $\beta = 90\%$ and for $\beta = 10\%$ in Figure 10 against ρ, and note again that the plots are always less than the claimed values 95% and 5%. This happens generally for all possible quantile levels β, that the actual Proportion is less than the alleged probability. It happens for any chosen value for the parameter; and it happens for any prior average of such θ values. If by contrast the center of curvature is to the right, then the actual Proportion is reversed and is larger than the alleged.

In summary, in the vector parameter context with a curved interest parameter the performance error with the Bayes calculation can be of asymptotic order $O(n^{-1/2})$.

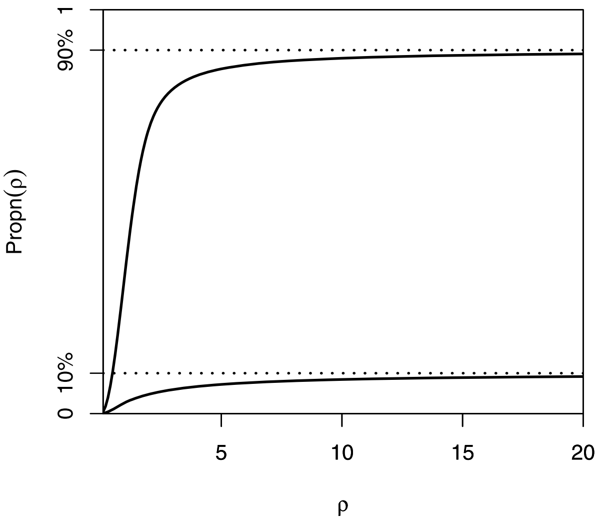


FIG. 10. *Proportion for claimed $\beta = 90\%$ and for claimed $\beta = 10\%$: strictly less than the claimed.*

6. NONLINEARITY AND MODEL CURVATURE: THE ERRORS ARE $O(n^{-1})$

(i) *The model and confidence bound.* Taylor series expansions provide a powerful means for examining the large sample form of a statistical model (see, e.g., Abebe et al., 1995; Andrews, Fraser and Wong, 2005; Cakmak et al., 1998). From such expansions we find that an asymptotic model to second order can be expressed as a location model and to third order can be expressed as a location model with an $O(n^{-1})$ adjustment that describes curvature.

Examples arise frequently in the vector parameter context. But for the scalar parameter context the common familiar models are location or scale models and thus without the curvature of interest here. A simple example with curvature, however, is the gamma distribution model: $f(y;\theta) = \Gamma^{-1}(\theta) y^{\theta-1} \exp\{-y\}$.

To illustrate the moderate curvature, we will take a very simple example where y is Normal$\{\theta, \sigma^2(\theta)\}$ and $\sigma^2(\theta)$ depends just weakly on the mean θ, and then in asymptotic standardized form we would have

$$\sigma^2(\theta) = 1 + \gamma\theta^2/2n$$

in moderate deviations. The β-level quantile for this normal variable y is

$$\begin{aligned} y_\beta(\theta) &= \theta + \sigma(\theta) z_\beta \\ &= \theta + z_\beta(1 + \gamma\theta^2/2n)^{1/2} \qquad (3) \\ &= \theta + z_\beta(1 + \gamma\theta^2/4n) + O(n^{-3/2}). \end{aligned}$$

The confidence bound $\hat{\theta}_\beta$ with β confidence above can be obtained from the usual Fisher inversion of $y = \theta + z_\beta(1 + \gamma\theta^2/4n)$: we obtain

$$\begin{aligned} \theta &= y - z_\beta(1 + \gamma\theta^2/4n) + O(n^{-3/2}) \\ &= y - z_\beta\{1 + \gamma(y - z_\beta)^2/4n\} + O(n^{-3/2}). \end{aligned}$$

Thus, the β level lower confidence quantile to order $O(n^{-3/2})$ is

$$\hat{\theta}^C(y) = y - z_\beta\{1 + \gamma(y - z_\beta)^2/4n\}, \qquad (4)$$

where we add the label C for confidence to distinguish it from other bounds soon to be calculated. See Figure 11.

(ii) *From confidence to likelihood.* We are interested in examining posterior quantiles for the adjusted normal model and in this section work from the confidence

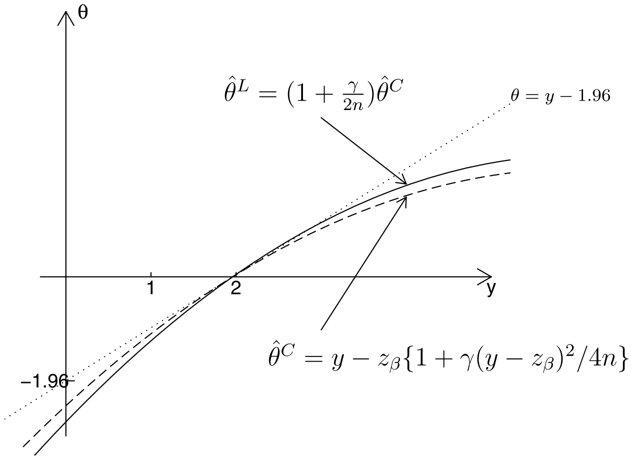


FIG. 11. *The* 97.5% *confidence quantile* $\hat\theta^C(y) = y - 1.96\{1 + \gamma(y - 1.96)^2/4n\}$. *The* 97.5% *likelihood quantile* $\hat\theta^L(y) = (1 + \frac{\gamma}{2n})[y - 1.96\{1 + \gamma(y - 1.96)^2/4n\}]$ *is a vertical rescaling about the origin*; *the* 97.5% *Bayes quantile* $\hat\theta^B(y)$ *with prior* $\exp\{a/n + c\theta/n\}$ *is a vertical rescaling plus a lift* a/n *and a tilt* cy/n. *Can this prior lead to a confidence presentation? No, unless the prior depends on the data or on the level* β.

quantile to the likelihood quantile, that is, to the posterior quantile with flat prior $\pi(\theta) = 1$; this route seems computationally easier than directly calculating a likelihood integral.

From Section 3 and formula (3) above, we have that the prior $\pi(\theta)$ that converts a likelihood $f^L(\theta) = L(\theta; y^0) = F_y(y^0; \theta)$ to confidence $f^C(\theta) = -F_{;\theta}(y^0; \theta)$ is

$$\begin{aligned}\frac{dy}{d\theta}\bigg|_{y^0} &= 1 + \gamma z\theta/2n|_{y^0}\\ &= 1 + \gamma(y^0 - \theta)\theta/2n + O(n^{-3/2})\\ &= \exp\{\gamma(y^0 - \theta)\theta/2n\} + O(n^{-3/2}).\end{aligned}$$

Then to convert in the reverse direction, from confidence $f^C(\theta)$ to likelihood $f^L(\theta)$, we need the inverse weight function

$$w(\theta) = \exp\{\gamma\theta(\theta - y^0)/2n\}. \tag{5}$$

Interestingly, this function is equal to 1 at $\theta = 0$ and at y^0, and is less than 1 between these points when $\gamma > 0$.

(iii) *From confidence quantile to likelihood quantile*. The weight function (5) that converts confidence to likelihood has the form $\exp\{a\theta/n^{1/2} + c\theta^2/2n\}$ with $a = -\gamma y^0/2n^{1/2}$ and $c = \gamma$. The effect of such a tilt and bending is recorded in the Appendix. The confidence quantile $\hat\theta^C_\beta$ given at (4) is a $1 - \beta$ quantile of the confidence distribution. Then using formula (10) in the Appendix, we obtain the formula for converting confidence quantile to likelihood quantile:

$$\begin{aligned}\hat\theta^L &= \hat\theta^C\left(1 + \frac{\gamma}{2n}\right) - \gamma y^0/2n + \gamma y^0/2n\\ &= \hat\theta^C\left(1 + \frac{\gamma}{2n}\right).\end{aligned} \tag{6}$$

Thus, the likelihood distribution is obtained from the confidence distribution by a simple scale factor $1 + \gamma/2n$; this directly records the consequence of the curvature added to the simple normal model by having $\sigma^2(\theta)$ depend weakly on θ.

(iv) *From likelihood quantile to posterior quantile*. Now consider a prior applied to the likelihood distribution. A prior can be expanded in terms of standardized coordinates and takes the form $\pi(\theta) = \exp(a\theta/n^{1/2} + c\theta^2/2n)$. The effect on quantiles is available from the Appendix and we see that a prior with tilt coefficient $a/n^{1/2}$ would excessively displace the quantile and thus would give posterior quantiles with bad behaving $\text{Propn}(\theta)$ in repetitions; accordingly, as a possible prior adjustment, we consider a tilt with just a coefficient a/n. We then examine the prior $\pi(\theta) = \exp(a\theta/n + c\theta^2/2n)$. First, we obtain the Bayes quantile in terms of the likelihood quantile as

$$\hat\theta^B = \hat\theta^L\left(1 + \frac{c}{2n}\right) + \frac{a}{n} + \frac{cy}{2n};$$

and then substituting for the likelihood quantile in terms of the confidence quantile (6) gives

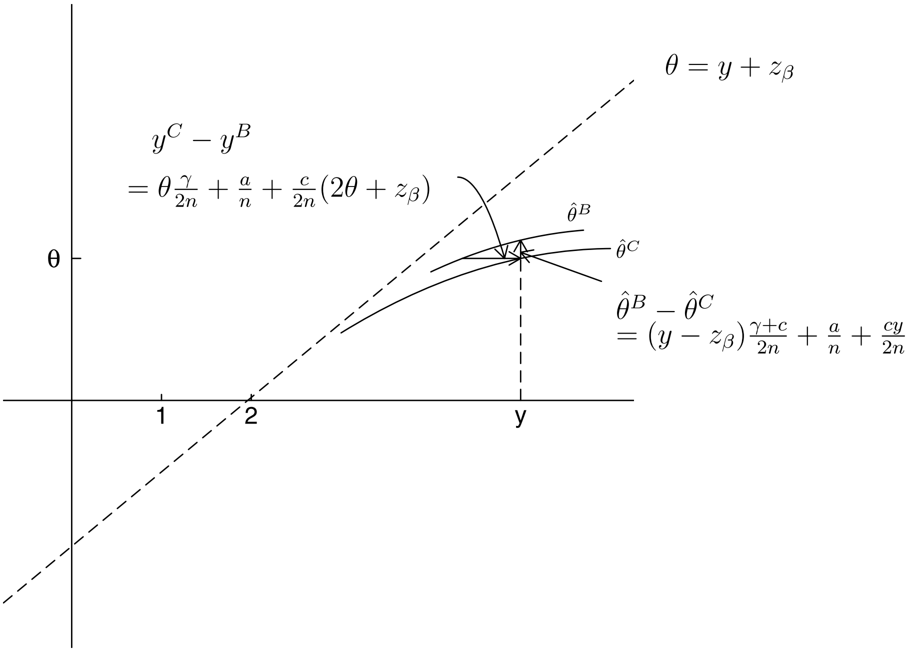
$$\hat\theta^B = \hat\theta^C\left(1 + \frac{\gamma + c}{2n}\right) + \frac{a}{n} + \frac{cy}{2n}. \tag{7}$$

For $\hat\theta^B(y)$ in (4) to be equal to $\hat\theta^C(y)$ in (7) we would need to have $c = -\gamma$ and then $a = \gamma y/2$. But this would give a data dependent prior. We noted the need for data dependent priors in Section 3, but we now have an explicit expression for the effect of priors on quantiles.

Now consider the difference in quantiles:

$$\begin{aligned}\hat\theta^B(y) - \hat\theta^C(y) &= \hat\theta^C\left(\frac{\gamma + c}{2n}\right) + \frac{a}{n} + \frac{cy}{2n}\\ &= (y - z_\beta)\frac{\gamma + c}{2n} + \frac{a}{n} + \frac{cy}{2n}\\ &= \frac{a}{n} + y\frac{\gamma + 2c}{2n} - z_\beta\frac{\gamma + c}{2n},\end{aligned}$$

where we have replaced $\hat\theta^C$ by $y - z_\beta$, to order $O(n^{-3/2})$; Figure 12 shows this difference as the vertical separation above a data value y. From the third



$\theta = y + z_\beta$
$y^C - y^B$
$= \theta\frac{\gamma}{2n} + \frac{a}{n} + \frac{c}{2n}(2\theta + z_\beta)$
$\hat\theta^B$
$\hat\theta^C$
θ
$\hat\theta^B - \hat\theta^C$
$= (y - z_\beta)\frac{\gamma+c}{2n} + \frac{a}{n} + \frac{cy}{2n}$
1
2
y

FIG. 12. *β-level quantiles. The difference $\hat\theta^B(y) - \hat\theta^C(y)$ is the vertical separation above y between quantile curves. The difference $y^C(\theta) - y^B(\theta)$ is the horizontal separation between curves as a function of θ.*

expression above we see that in the presence of model curvature γ the Bayesian quantile can achieve the quality of confidence only if the prior is data dependent or dependent on the level β.

Similarly, we can calculate the horizontal separation corresponding to a θ value, and obtain

$$\begin{aligned} y^C(\theta) - y^B(\theta) &= \theta\frac{\gamma + c}{2n} + \frac{a}{n} + \frac{c}{2n}(\theta + z_\beta) \\ &= \theta\frac{\gamma}{2n} + \frac{a}{n} + \frac{c}{2n}(2\theta + z_\beta). \end{aligned} \tag{8}$$

This gives the quantile difference, the confidence quantile less the Bayes quantile, as a function of θ; see Figure 12, and observe the horizontal separation to the right of a parameter value θ.

A Bayes quantile can not generate true statements concerning a parameter with the reliability of confidence unless the model curvature is zero, that is, unless the model is of the special location form where Bayes coincides confidence. The Bayes approach can thus be viewed as having a long history of misdirection.

Now let θ designate the true value of the parameter θ, and suppose we examine the performance of the Bayesian and frequentist posterior quantiles. In repetitions the actual proportion of instances where $y < y^C(\theta)$ is of course β. The actual proportion of cases with $y < y^B(\theta)$ is then

$$\mathrm{Propn}(\theta) = \beta - \left\{\theta\frac{\gamma}{2n} + \frac{a}{n} + \frac{c}{2n}(2\theta + z_\beta)\right\}\phi(z_\beta),$$

where for the terms of order $O(n^{-1})$ it suffices to use the $N(\theta, 1)$ distribution for y. The Bayes calculation claims the level β. The choice $a = 0, c = 0$ gives a flat prior in the neighborhood of $\theta = 0$ which is the central point of the model curvature. With such a choice the actual Proportion from the Bayes approach is deficient by the amount $\theta\gamma\phi(z_\beta)/2n$. For a claimed $\beta = 50\%$ quantile see Figure 13 for the actual Proportion and for a claimed $\beta = 90\%$ or $\beta = 10\%$ see Figure 14.

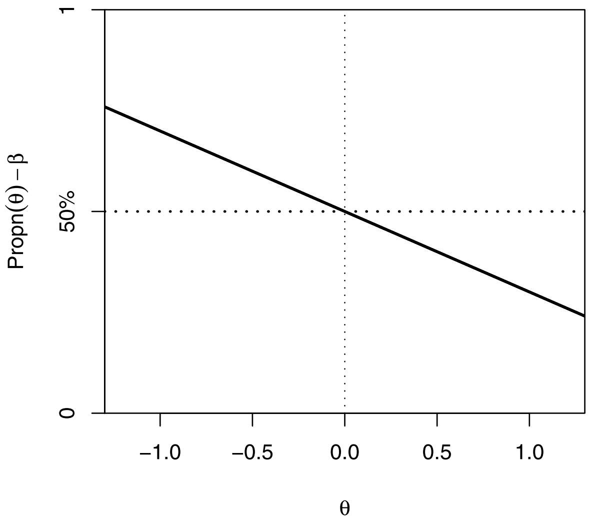


FIG. 13. *The actual Proportion with claimed level $\beta = 50\%$.*

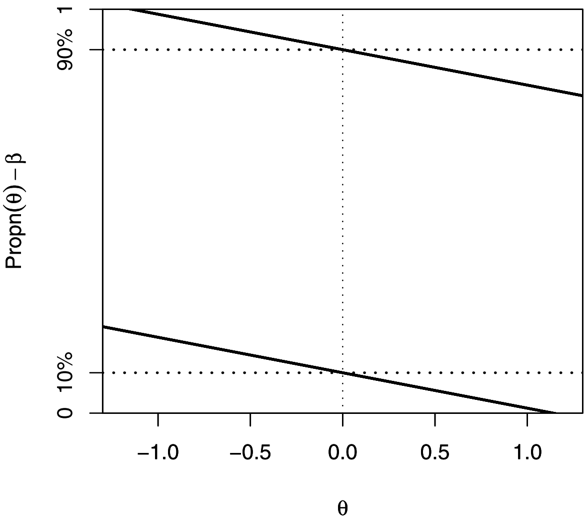


FIG. 14. *The actual Proportion with claimed levels $\beta = 90\%$ and $\beta = 10\%$.*

Thus, the β quantile by Bayes is consistently below the claimed level β for positive values of θ, and consistently above the claimed level for negative values of θ.

In summary, even in the scalar parameter context, an elementary departure from simple linearity can lead to a performance error for the Bayes calculation of asymptotic order $O(n^{-1})$. And, moreover, it is impossible by the Bayes method to duplicate the standard confidence bounds: a stunning revelation!

7. THE PARADIGM

The Bayes proposal makes critical use of the conditional probability formula $f(y_1|y_2^0) = cf(y_1, y_2^0)$. In typical applications the formula has variables y_1 and y_2 in a temporal order: the value of the first y_1 is inaccessible and the value of the second y_2 is observed with value, say, y_2^0. Of course, the value of the first y_1 has been realized, say, y_1^r, but is concealed and is unknown. Indeed, the view has been expressed that the only probabilities possible concerning such an unknown y_1^r are the values 0 or 1 and we don't know how they would apply to that y_1^r. We thus have the situation where there is an unknown constant y_1^r, a constant that arose antecedent in time to the observed value y_2^0, and we want to make probability statements concerning that unknown antecedent constant. As part of the temporal order we also have that the joint density became available in the order $f(y_1)$ for the first variable followed by $f(y_2|y_1)$ for the second; thus, $f(y_1, y_2^0) = f(y_1)f(y_2^0|y_1)$.

The conditional probability formula itself is very much part of the theory and practice of probability and statistics and is not in question. Of course, limit operations are needed when the condition $y_2 = y_2^0$ has probability zero leading to a conditional probability expression with a zero in the denominator, but this is largely technical.

A salient concern seemingly centers on how probabilities can reasonably be attached to a constant that is concealed from view? The clear answer is in terms of what *might* have occurred given the same observational information: the corresponding picture is of many repetitions from the joint distribution giving pairs (y_1, y_2); followed by selection of pairs that have exact or approximate agreement $y_2 = y_2^0$; and then followed by examining the pattern in the y_1 values among the selected pairs. The pattern records what would have occurred for y_1 among cases where $y_2 = y_2^0$; the probabilities arise both from the density $f(y_1)$ and from the density $f(y_2|y_1)$. Thus, the initial pattern $f(y_1)$ when restricted to instances where $y_2 = y_2^0$ becomes modified to the pattern $f(y_1|y_2^0) = cf(y_1, y_2^0) = cf(y_1)f(y_2^0|y_1)$.

Bayes (1763) promoted this conditional probability formula and its interpretation, for statistical contexts that had no preceding distribution for θ and he did so by introducing the mathematical prior. He did provide, however, a motivating analogy and the analogy did have something extra, an objective and real distribution for the parameter, one with probabilities that were well defined by translational invariance. Such a use of analogy in science is normally viewed as wrong, but the needs for productive methodology were high at that time.

If $\pi(\theta)$ is treated as being real and descriptive of how the value of the parameter arose in the application, it would follow that the preceding conditional probability analysis would give the conditional description

$$\begin{aligned}\pi(\theta|y^0) &= c\pi(\theta)f(y^0;\theta)\\ &= c\pi(\theta)L^0(\theta).\end{aligned}$$

The interpretation for this would be as follows: In many repetitions from $\pi(\theta)$, if each θ value was followed by a y from the model $f(y;\theta)$, and if the instances (θ, y) where y is close to y^0 are selected, then the pattern for the corresponding θ values would be $c\pi(\theta)L^0(\theta)$. In other words, the initial relative frequency $\pi(\theta)$ for θ values is modulated by $L^0(\theta)$ when we select using $y = y^0$; this gives the modulated frequency pattern $c\pi(\theta)L^0(\theta)$. The conditional probability formula as used in this context is often referred to as the Bayes formula or Bayes theorem, but as a probability formula it long predates Bayes and is generic;

for the present extended usage it is also referred to as the Bayes paradigm (Bernardo and Smith, 1994).

The Bayes' example as discussed in Sections 2 and 3 examined a location model $f(y-\theta)$ and the only prior that could represent location invariance is the constant or flat prior in the location parameterization, that is, $\pi(\theta)=c$. This of course does not satisfy the probability axioms, as the total probability would be ∞. The step, however, from just a set of θ values with related model invariance to a distribution for θ has had the large effect of emphasizing likelihood $L^0(\theta)$, as defined by Fisher (1935). And it has also had the effect, perhaps unwarranted, of suggesting that the mathematical posterior distribution obtained from the paradigm could be treated as a distribution of real probability. If the parameter to variable relationship is linear, then Section 3 shows that the calculated values have the confidence (Fisher, 1935; Neyman, 1937) interpretation. But if the relationship is nonlinear, then the calculated numbers can seriously fail to have that confidence property, as determined in Sections 4–6; and indeed fail to have anything with behavior resembling probability. The mathematical priors, the invariant priors and other generalizations are often referred to in the current Bayesian literature as objective priors, a term that is strongly misleading.

In other contexts, however, there may be a real source for the parameter θ, sources with a known distribution, and thus fully entitled to the term objective prior; of course, such examples do not need the Bayes approach, they are immediately analyzable by probability calculus. And, thus, to use objective to also refer to the mathematical priors is confusing.

In short, the paradigm does not produce probabilities from no probabilities. And if the required linearity for confidence is only approximate, then the confidence interpretation can correspondingly be just approximate. And in other cases even the confidence interpretation can be substantially unavailable. Thus, to claim probability when even confidence is not applicable does seem to be fully contrary to having acceptable meaning in the language of the discipline.

8. OPTIMALITY

Optimality is often cited as support for the Bayes approach: If we have a criterion of interest that provides an assessment of a statistical procedure, then optimality under the criterion is available using a procedure that is optimal under some prior average of the model. In other words, if you want optimality, it suffices to look for a procedure that is optimal for the prior-average version of the model. Thus, restrict one's attention to Bayes solutions and just find an appropriate prior to work from. It sounds persuasive and it is important.

Of course, a criterion as mentioned is just a numerical evaluation and optimality under one such criterion may not give optimality under some other criterion; so the choice of the criterion can be a major concern for the approach. For example, would we want to use the length of a posterior interval as the criterion or say the squared length of the interval or some other evaluation; it makes a difference because the optimality has to do with an average of values for the criterion and this can change with change in the criterion.

The optimality approach can lead to interesting results but can also lead to strange trade-offs; see, for example, Cox (1958) and Fraser and McDunnough (1980). For if the model splits with known probabilities into two or several components, then the optimality can create trade-offs between these; for example, if data sometimes is high precision and sometimes low precision and the probabilities for this are available, then the search for an optimum mean-length confidence interval at some chosen level can give longer intervals in the high precision cases and shorter intervals in the low precision cases as a trade-off toward optimality and toward intervals that are shorter on average. It does sound strange but the substance of this phenomenon is internal to almost all model-data contexts.

Even with a sensible criterion, however, and without the compound modeling and trade-offs just mentioned, there are serious difficulties for the optimality support for the Bayes approach. Consider further the example in Section 6 with a location Normal variable where the variance depends weakly on the mean: y is Normal$\{\theta,\sigma^2(\theta)\}$ with $\sigma^2(\theta)=1+\gamma\theta^2/2n$ and where we want a bound $\hat{\theta}_\beta(y)$ for the parameter θ with reliability β for the assertion that θ is larger than $\hat{\theta}_\beta(y)$.

From confidence theory we have immediately (4) that

$$\hat{\theta}(y)=\hat{\theta}^C(y)=y-z_\beta\{1+\gamma(y-z_\beta)^2/4n\}$$

with $O(n^{-3/2})$ accuracy in moderate deviations. What is available from the Bayes approach? A prior $\pi(\theta)=\exp\{a\theta/n^{1/2}+c\theta^2/2n\}$ gives the posterior bound

$$\hat{\theta}^\beta(y)=\hat{\theta}^C(y)\{1+c/2n\}+\frac{a}{n}+\frac{cy}{2n}.$$

The actual Proportion for the β level confidence bound is exactly β. The actual Proportion, however, for the Bayes bound as derived (8) is

$$\beta-\left\{\theta\frac{\gamma}{2n}+\frac{a}{n}+\frac{c}{2n}(2\theta+z_\beta)\right\}\phi(z_\beta);$$

and there is no choice for the prior, no choice for a and c, that will make the actual equal to the nominal unless the model has nonzero curvature γ.

We thus have that a choice of prior to weight the likelihood function can not produce a β level bound. But a β level bound is available immediately and routinely from confidence methods, which does use more than just the observed likelihood function.

Of course, in the pure location case the Bayes approach is linear and gives confidence. If there is nonlinearity, then the Bayes procedure can be seriously inaccurate.

9. DISCUSSION

Bayes (1763) introduced the observed likelihood function to general statistical usage. He also introduced the confidence distribution when the application was to the special case of a location model; the more general development (Fisher, 1930) came much later and the present name confidence was provided by Neyman (1937). Lindley (1958) then observed that the Bayes derivation and the Fisher (1930) derivation coincided only for location models; this prompted continuing discord as to the merits and validity of the two procedures in providing a probability-type assessment of an unknown parameter value.

A distribution for a parameter value immediately makes available a quantile for that parameter, at any percentage level of interest. This means that the merits of a procedure for evaluating a parameter can be assessed by examining whether the quantile relates to the parameter in anything like the asserted rate or level asserted for that quantile. The examples in Sections 4–6 demonstrate that departure from linearity in the relation between parameter and variable can seriously affect the ability of likelihood alone to provide reliable quantiles for the parameter of interest.

There is of course the question as to where the prior comes from and what is its validity? The prior could be just a device as with Bayes original proposal, to use the likelihood function directly to provide inference statements concerning the parameter. This has been our primary focus and such priors can reasonably be called default priors.

And then there is the other extreme where the prior describes the statistical source of the experimental unit or more directly the parameter value being considered. We have argued that these priors should be called objective and then whether to use them to perform the statistical analysis is a reasonable question.

Between these two extremes are many variations such as subjective priors that describe the personal views of an investigator and elicited priors that represent some blend of the background views of those close to a current investigation. Should such views be kept separate to be examined in parallel with objective views coming directly from the statistical investigation itself or should they be blended into the computational procedure applied to the likelihood function alone? There would seem to be strong arguments for keeping such information separate from the analysis of the model with data; any user could then combine the two as deemed appropriate in any subsequent usage of the information.

Linearity of parameters and its role in the Bayesian frequentist divergence is discussed in Fraser, Fraser and Fraser (2010a). Higher order likelihood methods for Bayesian and frequentist inference were surveyed in Bédard, Fraser and Wong (2007), and an original intent there was to include a comparison of the Bayesian and frequentist results. This, however, was not feasible, as the example used there for illustration was of the nice invariant type with the associated theoretical equality of common Bayesian and frequentist probabilities; thus, the anomalies discussed in this paper were not overtly available there.

10. SUMMARY

A probability formula was used by Bayes (1763) to combine a mathematical prior with a model plus data; it gave just a mathematical posterior, with no consequent objective properties. An analogy provided by Bayes did have a real and descriptive prior, but it was not part of the problem actually being examined.

A familiar Bayes example uses a special model, a location model; and the resulting intervals have attractive properties, as viewed by many in statistics.

Fisher (1935) and Neyman (1937) defined confidence. And the Bayes intervals in the location model case are seen to satisfy the confidence derivation, thus providing an explanation for the attractive properties.

The only source of variation available to support a Bayes posterior probability calculation is that provided by the model, which is what confidence uses.

Lindley (1958) examined the probability formula argument and the confidence argument and found that they generated the same result only in the Bayes location model case; he then judged the confidence argument to be wrong.

If the model, however, is not location and, thus, the variable is not linear with respect to the parameter, then a Bayes interval can produce correct answers at a rate quite different from that claimed by the Bayes probability calculation; thus, the Bayes posterior may be an unreliable presentation, an unreliable approximation to confidence, and can thus be judged as wrong.

The failure to make true assertions with a promised reliability can be extreme with the Bayes use of mathematical priors (Stainforth et al., 2007; Heinrich, 2006).

The claim of a probability status for a statement that can fail to be approximate confidence is misrepresentation. In other areas of science such false claims would be treated seriously.

Using weighted likelihood, however, can be a fruitful way to explore the information available from just a likelihood function. But the failure to have even a confidence interpretation deserves more than just gentle caution.

A personal or a subjective or an elicited prior may record useful background to be recorded in parallel with a confidence assessment. But to use them to do the analysis and just get approximate or biased confidence seems to overextend the excitement of exploratory procedures.

APPENDIX

Tilting, Bending and Quantiles

Consider a variable y that has a Normal$(\theta; 1)$ distribution and suppose that its density is subject to an exponential tilt and bending as described by the modulating factor $\exp\{ay + cy^2/2\}$. It follows easily by completing the square in the exponent that the new variable, say, $\tilde{y}$, is also normal but with mean $(\theta + a)/(1 - c)$ and variance $1/(1 - c)$. In particular, we can write

$$\tilde{y} = \frac{\theta + a}{1 - c} + (1 - c)^{-1/2} z,$$

where z is standard normal. And if we let z_β be the β quantile of the standard normal with $\beta = \Phi(z_\beta)$, then the β quantile of $\tilde{y}$ is

$$\tilde{y}_\beta = \frac{\theta + a}{1 - c} + (1 - c)^{-1/2} z_\beta.$$

Thus, with the Normal$(\theta, 1)$ we have that tilting and bending just produce a location scale adjustment to the initial variable.

Now suppose that $y = \theta + z$ is Normal$(\theta; 1)$ to third order, and suppose further that its density receives an exponential tilting and bending described by the factor $\exp\{ay/n^{1/2} + cy^2/2n\}$. Then from the preceding we have that the new variable can be expressed in terms of preceding variables as

$$\begin{aligned} \tilde{y} &= \frac{\theta + a/n^{1/2}}{1 - c/n} + (1 - c/n)^{-1/2} z \\ &= \theta(1 + c/n) + a/n^{1/2} + (1 + c/2n) z \qquad (9) \\ &= y(1 + c/2n) + a/n^{1/2} + \theta c/2n, \end{aligned}$$

where succeeding lines use adjustments that are $O(n^{-3/2})$. The second line on the right gives quantiles in terms of the standard normal and the third line gives quantiles in terms of the initial variable y.

One application for this arises with posterior distributions. Suppose that $\theta = y^0 + z$ is Normal$(y^0, 1)$ to third order and that its density receives a tilt and bending described by $\exp(a\theta/n^{1/2} + c\theta^2/2n)$. We then have from (9) that the modified variable can be expressed as

$$\begin{aligned} \tilde{\theta} &= y^0(1 + c/n) + a/n^{1/2} + (1 + c/2n) z \\ &= \theta(1 + c/2n) + a/n^{1/2} + y^0 c/2n, \end{aligned} \qquad (10)$$

to order $O(n^{-3/2})$.

ACKNOWLEDGMENTS

The author expresses deep appreciation to Nancy Reid for many helpful discussions from various viewpoints; and we join to express deep appreciation to Pekka Sinervo who introduced us to statistical concerns in High Energy Physics and then took valuable time from Decanal duties to clarify contextual issues. Also very special thanks go to Ye Sun, Tara Cai and Kexin Ji for many contributions toward this research, including numerical computation and preparation of figures and work on the manuscript. The Natural Sciences and Engineering Research Council of Canada has provided financial support. The author also thanks the reviewers for very helpful comments.

REFERENCES

ABEBE, F., CAKMAK, S., CHEAH, P. K., FRASER, D. A. S., KUHN, J., MCDUNNOUGH, P., REID, N. and TAPIA, A. (1995). Third order asymptotic model: Exponential and location type approximations. *Parisankhyan Samikkha* **2** 25–33. MR1613190

ANDREWS, D. F., FRASER, D. A. S. and WONG, A. C. M. (2005). Computation of distribution functions from likelihood information near observed data. *J. Statist. Plann. Inference* **134** 180–193. MR2146092

BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* **53** 370–418; **54** 296–325. Reprinted in *Biometrika* **45** (1958) 293–315.

BÉDARD, M., FRASER, D. A. S. and WONG, A. (2007). Higher accuracy for Bayesian and frequentist inference: Large sample theory for small sample likelihood. *Statist. Sci.* **22** 301–321. MR2416807

BERNARDO, J.-M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147. MR0547240

BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester. MR1274699

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

CAKMAK, S., FRASER, D. A. S., MCDUNNOUGH, P., REID, N. and YUAN, X. (1998). Likelihood centered asymptotic model: Exponential and location model versions. *J. Statist. Plann. Inference* **66** 211–222. MR1614476

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372. MR0094890

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 189–233. MR0365805

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **222** 309–368.

FISHER, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **26** 528–535.

FISHER, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics B* 391–398.

FRASER, A. M. FRASER, D. A. S. and FRASER, M. J. (2010a). Parameter curvature revisited and the Bayesian frequentist divergence. *J. Statist. Res.* **44** 335–346.

FRASER, A. M., FRASER, D. A. S. and STAICU, A. M. (2010c). Second order ancillary: A differential view with continuity. *Bernoulli* **16** 1208–1223.

FRASER, D. A. S. and MCDUNNOUGH, P. (1980). Some remarks on conditional and unconditional inference for location-scale models. *Statist. Hefte* (*N.F.*) **21** 224–231. MR0639293

FRASER, D. A. S., REID, N. and WONG, A. (2004). Setting confidence intervals for bounded parameters: A different perspective. *Phys. Rev. D* **69** 033002.

FRASER, D. A. S., REID, N., MARRAS, E. and YI, G. Y. (2010b). Default prior for Bayesian and frequentist inference. *J. Roy. Statist. Soc. Ser. B* **75** 631–654.

HEINRICH, J. (2006). The Bayesian approach to setting limits: What to avoid? In *Statistical Problems in Particle Physics, Astrophysics and Cosmology* (L. Lyons and Ü. M. Karagöz, eds.) 98–102. Imperial College Press, London. MR2270215

JEFFREYS, H. (1939). *Theory of Probability*, 3rd ed. Oxford Univ. Press, Oxford.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London. Ser. A* **186** 453–461. MR0017504

LINDLEY, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. Ser. B* **20** 102–107. MR0095550

MANDELKERN, M. (2002). Setting confidence intervals for bounded parameters. *Statist. Sci.* **17** 149–172. MR1939335

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **237** 333–380.

REID, N. and FRASER, D. A. S. (2003). Likelihood inference in the presence of nuisance parameters. In *Proceedings of PHYSTAT2003* (L. Lyons, R. Mount and R. Reitmeyer, eds.) 265–271. SLAC E-Conf C030908.

STAINFORTH, D. A., ALLEN, M. R., TREDGER, E. R. and SMITH, L. A. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 2145–2162. See also: Gambling on tomorrow. Modelling the Earth's climate mathematically is hard already. Now a new difficulty is emerging. *Economist* **August 18** (2007) 69.

WASSERMAN, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 159–180. MR1747402

WOODROOFE, M. and WANG, H. (2000). The problem of low counts in a signal plus noise model. *Ann. Statist.* **28** 1561–1569. MR1835031

ZHANG, T. and WOODROOFE, M. (2003). Credible and confidence sets for restricted parameter spaces. *J. Statist. Plann. Inference* **115** 479–490. MR1985880