A RAND NOTE

Assessing the Accuracy of Normal Approximations

James S. Hodges

August 1988

40 Years 1948-1988

RAND

This research was supported by The RAND Corporation as part of its program of public service.

Published by The RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138

A RAND NOTE

N-2794-RC

Assessing the Accuracy of Normal Approximations

James S. Hodges

August 1988

40 Years 1948-1988

RAND

Assessing the Accuracy of Normal Approximations

JAMES S. HODGES*

The widespread applicability and use of normal approximations creates a need for methods for assessing their accuracy in an operational fashion. In this article, two new methods are proposed and exemplified. These new methods are based on a comparison of the level curves of an exact likelihood with the level curves of the usual normal approximation to it and on a comparison of selected line integrals of an exact density and a normal approximation to it. The operational usefulness of these methods is compared with the operational usefulness of two well-known existing approaches to assessing the accuracy of approximations, namely convergence rates and the curvature methods of Bates and Watts (1980).

As an example, the second method is applied to the Lee–Geisser (1972, 1975) approximation to the predictive distribution of dental measurements on girls at the ages of 8, 10, 12, and 14 years. In this application, the method shows that the 50% central (predictive) probability region has actual (predictive) probability content between 3% and 6% higher, that the approximate 80% central region has actual probability content between 1.7% lower and 4.4% higher, and that the approximate 95% central region has actual probability content between 3.7% lower and 5.6% higher.

The first method gives the smallest approximate likelihood region that contains the exact likelihood region and the largest approximate likelihood region contained in the exact region.

KEY WORDS: Confidence region; Curvature methods; Likelihood region; Posterior density; Predictive density.

1. INTRODUCTION

Approximations based on the normal distribution are ubiquitous in statistics. Practitioners in all foundational schools use them: Bayesians use normal approximations for posterior densities (Zellner 1971, p. 32); the likelihood can often be approximated by a function proportional to a normal density; in large samples the distribution of the maximum likelihood estimate is approximately normal (Cox and Hinkley 1974, p. 294); and level curves of the likelihood (i.e., likelihood regions), which are approximate confidence regions, can in turn be approximated by ellipsoids, which is equivalent to using a normal approximation to the likelihood.

With this widespread use comes a need to assess how well these approximations work, in some operational way. One such operational approach is through Fisher's disjunction (1973, sec. III-1). The disjunction is as follows: if an unexpected result is observed, then *either* something improbable has happened *or* the postulated probability mechanism is not correct. This can be restated with a more data-analytic flavor as follows: if some apparent feature of the data strongly contradicts your expectation, then *either* something improbable has happened *or* one or more of the assumptions that led you to look at that feature of the data with a particular expectation is incorrect. One of these latter assumptions is that any approximations used were accurate enough in the instance in question not to have caused the observed but unexpected result. To be useful, then, a method for assessing the accuracy of an approximation should tell its user whether or not the approximate entity he has calculated displays a particular feature simply because the approximation is inaccurate.

In Section 2 I present two new methods for assessing the accuracy of approximations, which are operational in this sense. In Section 3 they are applied to two examples, and in Section 4 they are compared with two well-known approaches to assessing the accuracy of approximations.

The two methods in Section 2 are defined without making approximations, but to compute them, some approximations must be made. It might appear that I have created a distinction without a difference: other accuracy assessment methods use approximations in defining their measures and can be computed exactly; my measures are defined without approximation, but approximations are needed to compute them. The difference is that improving on any given implementation of my methods simply requires a better computing technique, but improving on the other methods would require an entirely new method.

2. TWO METHODS FOR ASSESSING THE ACCURACY OF NORMAL APPROXIMATIONS

The first of these methods is applicable to comparing an exact probability density with a normal approximation, and the second is applicable to comparing an exact likelihood with a normal approximation. The idea is to find functionals of the functions to be compared that are both informative and practical. The first accuracy assessment method is based on comparing selected line integrals of the densities, and the second is based on a comparison of level curves of the likelihood and its approximation.

2.1 Integrals Method for Comparing Densities

Let g be a unimodal nonzero probability density on real p-space R^p, and let f be the density of a p-dimensional $N(\mu, \Sigma)$ random variable with Σ nonsingular. Without loss of generality, let $\mu = 0$. Now if $f(x) = g(x)$ for all x, then for any function $l(x)$ and any curve segment S, the line integral $\int_S l(x)g(x)\,dx$ will be close to $\int_S l(x)f(x)\,dx$. The function l and curve segment S that I propose yield

$$I(f) = K_1 \int_a^b r^{p-1} f(r\Sigma^{1/2}h)\,dr,$$

where $K_1 = 2\pi^{p/2}|\Sigma|^{1/2}/\Gamma(p/2)$, for h a unit length direction vector and a and b supplied by the user, and

$$I(g, h) = K_1 \int_a^b r^{p-1} g(r\Sigma^{1/2}h)\,dr \qquad (2.1)$$

* James S. Hodges is Associate Statistician, Department of Economics and Statistics, The RAND Corporation, Santa Monica, CA 90406. This work was supported by National Institutes of Health Grant GM 25271, by a University of Minnesota Graduate School Doctoral Dissertation Fellowship, by a grant from the University Computer Center of the University of Minnesota, and by RAND Corporation funds. The author thanks Seymour Geisser, Dennis Cook, Luke Tierney, Robert McCulloch, Maureen Lahiff, and the referees for helpful comments.

Reprinted from the Journal of the American Statistical Association, March 1987, Vol. 82, No. 397, pp.149–154,

[the dependence of $I(f)$ and $I(g, h)$ on a and b will be suppressed for convenience]. The motivation of this choice follows.

Let the scalars $b > a \geq 0$ be given, and define the ellipsoids $E_b = \{x \mid x'\Sigma^{-1}x \leq b^2\} = \{x \mid x = r\Sigma^{1/2}h, \|h\| = 1, r \in [0, b]\}$ and E_a. Then

$$\Pr_f(E_b - E_a) = \int_{x \in E_b - E_a} f(x)\, dx$$

$$= |\Sigma|^{1/2} \int_{a^2 < y'y \leq b^2} f(\Sigma^{1/2}y)\, dy,$$

for $y = \Sigma^{-1/2}x$. By transforming to polar coordinates (e.g., Mardia, Kent, and Bibby 1979, pp. 35–36), where $r = (y'y)^{1/2}$ and h is a direction vector, it follows that

$$\Pr_f(E_b - E_a) = K_1 \int_a^b r^{p-1} f(r\Sigma^{1/2}h)\, dr = I(f),$$

for h an arbitrary direction vector. Define $I(g, h)$ analogously as in (2.1), for an arbitrary direction vector h. For reasons that I will explain presently, the suggested usage for $I(g, h)$ is to find h^* and h_* that maximize and minimize, respectively, $I(g, h)$ and to examine $I(g, h)/I(f)$ for those values of h.

To interpret $I(g, h^*)$, define a function g^* as follows. For each fixed value of r in the half-open interval $(a, b]$, with h varying over the unit length direction vectors, define $g^*(r\Sigma^{1/2}h) = g(r\Sigma^{1/2}h^*)$. For $r \leq a$ or $r > b$, define g^* arbitrarily. Then for $a < r \leq b$, g^* has ellipsoidal contours, and by changing to polar coordinates as previously mentioned, it is easy to show that

$$\int_{x \in E_b - E_a} g^*(x)\, dx = K_1 \int_a^b r^{p-1} g(r\Sigma^{1/2}h^*)\, dr = I(g, h^*).$$

That is, if g^* is formed with the same shape in all directions (expanded and contracted by $\Sigma^{1/2}$) as g has in the direction $\Sigma^{1/2}h^*$, then $I(g, h^*)$ is the area under g^* over the region $E_b - E_a$. By an exactly analogous argument, if g_* is formed having the same shape in all directions (expanded and contracted by $\Sigma^{1/2}$) as g has in the direction $\Sigma^{1/2}h_*$, then $I(g, h_*)$ is the area under g_* over the region $E_b - E_a$.

Further, $I(g, h_*) \leq \Pr_g(E_b - E_a) \leq I(g, h^*)$; this follows easily by an argument similar to that noted previously. Using $I(g, h^*)$ and $I(g, h_*)$, then, amounts to using the shape of g along rays from f's mode μ in two directions, $\Sigma^{1/2}h^*/\|\Sigma^{1/2}h^*\|$ and $\Sigma^{1/2}h_*/\|\Sigma^{1/2}h_*\|$, to construct two surfaces g^* and g_* that have ellipsoidal contours and that satisfy

$$\int g^*(x)\, dx \geq \int g(x)\, dx \geq \int g_*(x)\, dx,$$

where all integrals are for $x \in E_b - E_a$ and the first and last integrals are maximized and minimized, respectively, subject to the construction. This latter inequality can be used to put bounds on the exact probability of regions $E_b - E_a$, in particular of approximate central regions of probability $1 - \alpha$ constructed by using approximations to posterior or predictive densities. That is, $100 \times I(g, h^*)/I(f)$ and $100 \times I(g, h_*)/I(f)$ are upper and lower bounds on the fraction of the approximate probability of the region that the exact probability attains. Judging whether these bounds indicate a large or small difference is discussed in the context of the example in Section 3.1.

To recapitulate, the "integrals" method of comparison, as this method will be called, consists of searching the directions h for those in which the integral

$$\int_a^b r^{p-1} g(\mu + r\Sigma^{1/2}h)\, dr$$

is large or small compared with

$$\int_a^b r^{p-1} f(\mu + r\Sigma^{1/2}h)\, dr.$$

These integrals are multiplied by the constant K_1 to permit the latter to be interpreted as a probability. Note that although $I(f)$ is a probability, $I(g, h^*)$ and $I(g, h_*)$ are not, and $I(g, h^*)$ can exceed unity.

2.2 Contours Method for Comparing Likelihoods

Suppose that $L(\theta; y)$ is the likelihood function of a p-dimensional parameter θ, where y is data. Under regularity conditions, if θ^0 is the true value of θ and $\hat{\theta}$ is the maximum likelihood estimator (MLE),

$$W(\theta^0) = -2[\log L(\theta^0; y) - \log L(\hat{\theta}; y)]$$

is approximately distributed as $\chi^2(p)$, a chi-squared random variable with p degrees of freedom (see Cox and Hinkley 1974, p. 322), so $W(\theta)$ is an approximate pivotal. An approximate $100(1 - \alpha)\%$ confidence region would be the region in parameter space satisfying

$$\{\theta \mid W(\theta) \leq c\} = \{\theta \mid g(\theta) \geq K_2\} \tag{2.2}$$

for

$$K_2 = |I_{\hat{\theta}}|^{1/2}(2\pi)^{-p/2} \exp(-c/2),$$

$$g(\theta) = K_2 \exp(c/2) L(\theta; y)/L(\hat{\theta}; y),$$

$$I_{\hat{\theta}} = ((-\partial^2 \log L(\theta; y)/\partial\theta_i \partial\theta_j|_{\theta=\hat{\theta}}))_{p \times p}$$

the observed Fisher information, and c satisfying $\Pr(\chi^2(p) \leq c) = 1 - \alpha$. This gives the level curves of the likelihood an approximate confidence interpretation.

Finding the aforementioned confidence region can be difficult, so $\log L(\theta; y)$ is commonly approximated by

$$\log L(\theta; y) \approx \log L(\hat{\theta}; y) - (\theta - \hat{\theta})' I_{\hat{\theta}} (\theta - \hat{\theta})/2.$$

Thus the region (2.2) can be approximated by

$$\{\theta \mid (\theta - \hat{\theta})' I_{\hat{\theta}} (\theta - \hat{\theta}) \leq c\} = \{\theta \mid f(\theta) \geq K_2\}, \tag{2.3}$$

where $f(\theta)$ is the density of Θ, an $N(\hat{\theta}, I_{\hat{\theta}}^{-1})$ random variable. Note that the nominal confidence coefficient $1 - \alpha$ is exactly

$$\Pr_f((\Theta - \hat{\theta})' I_{\hat{\theta}} (\Theta - \hat{\theta}) \leq c) = \Pr_f(f(\Theta) \geq K_2)$$

$$= \Pr(\chi^2(p) \leq c).$$

(This manipulation with f is purely formal; no fiducial

probability ideas are being invoked.) Now if θ_0 is a point on the contour $g(\theta) = k$ of g, the contour of f that passes through θ_0 is $f(\theta) = f(\theta_0)$ and, following the previous calculations, this contour is the normal approximate confidence region with nominal confidence coefficient

$$\Pr_f(f(\Theta) \geq f(\theta_0)) = \Pr(\chi^2(p) \leq -2 \log[K_3 f(\theta_0)]), \quad (2.4)$$

where $K_3 = (2\pi)^{p/2}|I_\theta|^{-1/2}$. Thus, for any given point on the contour (2.2), we can compute the nominal confidence coefficient of the approximate region (2.3) passing through that point. If we find a point on the exact contour (2.2) such that the nominal confidence coefficient of the approximate region passing through that point is much larger or smaller than $1 - \alpha$, we have evidence of a deficiency in the approximation.

The "contours" method, as this method will be called, consists of finding the point on the exact contour (2.2) such that the corresponding nominal confidence coefficient (2.4) of the approximate region passing through that point is maximized, and the point on the exact contour such that the corresponding nominal confidence coefficient of the approximate region passing through that point is minimized. The resulting nominal confidence coefficients are, respectively, the nominal confidence coefficient of the smallest approximate region that contains the exact region and the nominal confidence coefficient of the largest approximate region contained in the exact region. This is a refinement of the method of Minkin (1983), who found an upper bound for the maximum nominal confidence coefficient. Judging whether the maximum and minimum nominal confidence coefficients differ substantially is discussed in the context of the example in Section 3.2.

To understand the contours method, let $f^c(\theta) = (\theta - \hat{\theta})' I_\theta (\theta - \hat{\theta})$ and $g^c = -2[\log L(\theta; y) - \log L(\hat{\theta}; y)]$ so that the approximate and exact likelihood regions are $\{\theta \mid f^c(\theta) \leq c\}$ and $\{\theta \mid g^c(\theta) \leq c\}$, respectively. The boundaries of the two regions are θ satisfying $c = f^c(\theta) = (\theta - \hat{\theta})' I_\theta (\theta - \hat{\theta})$ and $c = g^c(\theta) = (\theta - \hat{\theta})' I_\theta (\theta - \hat{\theta}) +$ {higher-order terms in the Taylor series expansion}. If the approximate region differs from the exact region, to the first order of approximation the difference is caused by the cubic terms in the expansion of g^c. Jennings (1986) devised a measure of the size of the cubic terms to characterize the adequacy of the normal approximation to the likelihood region for logistic regression models; the contours method can be understood as a refinement of Jennings's method.

Often the observed information matrix I_θ in (2.2) and (2.3) is replaced by the expected Fisher information

$$I^e = ((E[-\partial^2 \log L(\theta; y)/\partial\theta_i \partial\theta_j]|_{\theta=\hat{\theta}}))_{p \times p}.$$

The development in this section follows without other alteration if I_θ is replaced by I^e.

The contours method can be defined more generally, to be used in comparing a function g with another function f satisfying the symmetry condition $f(\mu + th) = f(\mu - th)$, for all real numbers t and unit length direction vectors h, where μ is the mode of f. In these more general cases, however, its interpretation is not as natural.

2.3 Computing the Assessment Methods

In computing the assessment methods, I assumed that g is the exact function (density or likelihood), and that f, an approximation to g, is an $N(\mu, \Sigma)$ density, for Σ nonsingular. Without loss of generality, μ can be taken to be the origin and Σ the identity matrix. In practice, it is necessary to transform the densities so that f is a standard normal density, evaluate the accuracy assessment measures, and then backtransform the directions that maximize and minimize the measures.

With these assumptions, it is easy to show that for the contours method, maximizing and minimizing the corresponding nominal confidence coefficients (2.4) are equivalent to finding the directions h in which the contour $g(x) = k$ is farthest from the origin and nearest to the origin, respectively. The objective function used in computing the contours method, then, is $S_C(h) = r_h$, where r_h solves

$$g(r_h h) = k. \quad (2.5)$$

Evaluating $S_C(h)$ can be a source of difficulty, because depending on the shape of g and the location of its mode, (2.5) can have any number of solutions. Any algorithm that evaluates S_C for use in computing the contours method will embody implicit assumptions about g's properties, and these constitute the approximation (noted at the end of Sec. 1) that is made in computing the contours method.

For the integrals method, I approximated

$$I(g, h) = (2\pi^{p/2}/\Gamma(p/2)) \int_a^b r^{p-1} g(rh)\, dr$$

by the trapezoidal approximation

$$I(g, h) \approx S_I(h)$$
$$= \pi^{p/2}(b - a)\{b^{p-1} g(bh) + a^{p-1} g(ah)\}/\Gamma(p/2)$$

in searching for h^* and h_*. In computing the integral for $I(g, h^*)$ and $I(g, h_*)$, I use a one-dimensional numerical integral. Clearly, it would be possible to use a finer approximation to $I(g, h)$, for example, to use an exponential curve instead of a trapezoid, but this single trapezoid has produced good answers in my test cases, even for densities with very unusual contours, and it runs quickly.

3. EXAMPLES

In this section, I give examples of the application of the two methods and discuss the issue of whether a detected difference between the approximation and the exact function is large.

3.1 Integrals Method

Geisser (1970) and Lee and Geisser (1972) developed a Bayesian analysis of growth curve models with particular interest in predictions from those models, and Lee and Geisser (1975) applied that analysis to the Potthoff–Roy data [given in Lee and Geisser (1975)]. In these models the $p \times N$ matrix of observables Y is assumed to have expected value $X\tau A$, where X is a known $p \times m$ matrix of rank $m < p$, A is a known $r \times N$ matrix of rank $r <$

N, and τ is an unknown $m \times r$ matrix. The columns of Y are assumed to be independent p-dimensional multivariate normal random variables with common unknown covariance matrix Σ. For the Potthoff–Roy data, Lee and Geisser assumed that Σ had the simple form $\Sigma = X\Gamma X' + Z\theta Z'$, for Z a known $p \times (p - m)$ matrix of rank $(p - m)$ satisfying $Z'X = 0$ and Γ and θ arbitrary $m \times m$ and $(p - m) \times (p - m)$ positive semidefinite matrices. This is called "Rao's simple structure," and a likelihood ratio test does not suggest any reason not to use it here.

One prediction of interest is of K future p-dimensional vectors forming the $p \times K$ matrix V, which has expectation $X\tau F$ for some appropriate known $r \times K$ matrix F. If the prior density $\pi(\Gamma^{-1}, \theta^{-1}, \tau) \propto |\Gamma|^{(m+1)/2}|\theta|^{(p-m+1)/2}$ is used for $(\Gamma^{-1}, \theta^{-1}, \tau)$, the exact predictive density for V can be derived, but it is awkward to evaluate. Its expectation and covariance matrix can be obtained, however, and Lee and Geisser (1972) proposed approximating the exact predictive density of V by a matrix normal density with expectation and covariance equal to those of the exact density. The accuracy of this approximation can be checked by using the integrals method; this will be exemplified with the Potthoff–Roy data.

These data consist of four dental measurements on each of 11 girls and 16 boys, taken at ages 8, 10, 12, and 14. The postulated growth curve is linear, so $p = 4$, $m = r = 2$, $N = 27$; the design matrices X and A are

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \end{bmatrix},$$

$$A = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}$$

(where the first block of A has 11 columns and the second block has 16 columns), and τ is 2×2. If we wish to predict the four measurements for the next girl, $K = 1$ and $F = (1, 0)'$. Then, from Geisser (1970), the exact predictive density of V is

$$g(V) = c_1|DYY'D'|^{27/2}|BSB'|^{25/2} \times |DYY'D' + DVV'D'|^{-14} \times |BSB' + c_2(BV - T)(BV - T)'|^{-13},$$

where $c_1 = 143/80\pi^2$, $c_2 = 11/12$, $T' = (22.648, .47955)$,

$$B = \frac{1}{20}\begin{bmatrix} 5 & 5 & 5 & 5 \\ -3 & -1 & 1 & 3 \end{bmatrix},$$

$$D = \frac{1}{20}\begin{bmatrix} 5 & -5 & -5 & 5 \\ -1 & 3 & -3 & 1 \end{bmatrix} \quad \text{for}$$

$$Z' = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{bmatrix},$$

$$DYY'D' = \begin{bmatrix} 7.1719 & -.32812 \\ -.32812 & 3.1994 \end{bmatrix}$$

with determinant 22.838, and

$$BSB' = \begin{bmatrix} 94.491 & 1.7008 \\ 1.7008 & 2.9586 \end{bmatrix}$$

with determinant 276.67 [interested readers are referred to Geisser (1970) for details].

From Lee and Geisser [1972, expression (6.4)], the exact predictive covariance of V is $\Sigma = 6(XBSB'X')/121 + ZDYY'D'Z'/24$, that is,

$$\Sigma = \begin{bmatrix} 5.9594 & 4.0349 & 4.2052 & 3.5307 \\ & 6.2242 & 3.6379 & 4.4879 \\ & & 6.4175 & 4.8190 \\ & & & 6.9167 \end{bmatrix},$$

and the exact predictive mean is $\mu = XTF = (21.209, 22.168, 23.127, 24.086)'$. This gives the approximate density

$$f(V) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-(V - \mu)'\Sigma^{-1}(V - \mu)/2\}.$$

Suppose that we want central regions of predictive probability p. For the normal approximation, these correspond to ellipsoids centered at the mode of the predictive distribution, μ. The integrals method of Section 2.1 can be used to compare the actual probability content of the approximate region with the nominal probability content, with $a = 0$ and b set so that $\Pr_f(E_b) = p$. I did this for $p = .5$, .8, and .95 and computed $I(g, h^*)/I(f)$ and $I(g, h_*)/I(f)$. The results are given in Table 1.

Table 1 has a simple interpretation. The actual probability content of the nominal 50% predictive region is between 3% and 6% higher; the actual probability content of the 80% region is between 1.7% lower and 4.4% higher; and the actual probability content of the 95% region is between 3.7% lower and 5.6% higher.

Are these deviations "large" or "small"? Statisticians have become accustomed by training and by the incautious use of work such as Bates and Watts (1980) to demand precise answers in the form of critical values: greater than some number c_u is too large, less than some number c_l is too small. In the assessment of an approximation's accuracy, this thinking is mistaken. The person using an approximation has a level of precision he needs or can live with or has a budget of effort and money of which some can be used to reduce approximation error in competition with other needs, and information about the accuracy of an approximation should be in a form suitable for addressing these kinds of concerns. I submit that the integrals method, as exemplified here, is in this form and that no further guidance about "critical values" is useful.

The information in Table 1 is a drastic condensation of the information available from doing the searches needed to execute the integrals method. My implementation of the method produces a local maximum or minimum for each starting value submitted and the direction vectors h for which those maxima and minima were obtained. This

Table 1. Results of Applying Integrals Method to the Lee–Geisser Approximation

p	$I(g, h_*)/I(f)$	$I(g, h^*)/I(f)$
.5	1.030	1.060
.8	.983	1.044
.95	.963	1.056

can be used (see Hodges 1985a, sec. 4.3, or Hodges 1985b, sec. 3.5b) to make a detailed comparison of different approximations. In higher-dimensional problems, some reduced representation of the exact and approximate densities is necessary, and the integrals method in this type of implementation provides one candidate.

3.2 Contours Method

In this section I apply the contours method to the Fieller–Creasy problem (see Cook and Witmer 1985). This example is somewhat contrived, but I include it because of its recent use by Cook and Witmer (1985) to illustrate a weakness in the accuracy assessment method of Bates and Watts (1980). By way of comparison, I first give a brief description of the methods of Bates and Watts and then give the example.

The approach of Bates and Watts (1980) is applicable to models with an observable $y = F(x_i; \theta) + \varepsilon_i$, where F is a known function of arguments x_i and θ, x_i is a known q-vector of observable regressors, θ is an unknown, unobservable p-vector of parameters, and the errors ε_i are iid $N(0, \sigma^2)$ for $\sigma^2 > 0$. For functions $F(x; \theta)$ that are nonlinear in θ—that is, nonlinear regression models—tractable exact frequentist results are available only for special cases (e.g., Gallant 1975). Because the added difficulty coincides with the nonlinearity of $F(x; \theta)$ as a function of θ, Bates and Watts's method depends on examination of local curvature properties of the solution locus $\eta(\theta) \equiv (F(x_1; \theta), \ldots, F(x_n; \theta))'$ around the MLE, which is also the least squares estimator for this model.

To define Bates and Watts's two curvature measures, define the function $\eta_h(t) = \eta(\hat{\theta} + th)$, where $\hat{\theta}$ is the MLE, t is a real number, and h is an arbitrary p-dimensional unit length direction vector. The collection of tangent vectors $\{\dot{\eta}_h = d\eta_h/dt|_{t=0} \mid \|h\| = 1\}$ forms the tangent plane of the solution locus at $\hat{\theta}$. The acceleration vector $\ddot{\eta}_h = d^2\eta_h(t)/dt^2|_{t=0}$ can be decomposed as $\ddot{\eta}_h^N + \ddot{\eta}_h^T$, where $\ddot{\eta}_h^N$ is the rate of change of $\dot{\eta}_h$ orthogonal to the tangent plane and $\ddot{\eta}_h^T$ is the rate of change of $\dot{\eta}_h$ parallel to the tangent plane. These two components are converted to curvatures by norming them and dividing by the square of the norm of $\dot{\eta}_h$ and then are rendered invariant to scale changes in y by multiplying them by s, where s^2 is the usual unbiased estimate of σ^2. These standardized curvatures are then maximized with respect to the direction h to give the maximum relative intrinsic curvature, corresponding to $\ddot{\eta}_h^N$, and the maximum relative parameter effects curvature, corresponding to $\ddot{\eta}_h^T$. The former measures curvature of the solution locus and is invariant under changes of parameterization. The latter measures curvature of $\eta_h(t)$ within the solution locus and depends on the parameterization. Bates and Watts gave a guide value with which these two curvatures should be compared. If either curvature is larger than the guide value, this indicates a problem with the usual normal approximation to the likelihood regions (equivalent to approximating the solution locus by a linear Taylor series); otherwise, no problem is indicated. Interested readers are referred to Bates and Watts (1980).

We can now apply the contours method to nonlinear regression models and compare it with the Bates and Watts measures. For ease of presentation, assume that σ^2 is known. In nonlinear regression problems, the exact likelihood region is commonly approximated by (2.3), with I^e used as the information matrix. This region is $\{\theta \mid (\theta - \hat{\theta})'V'V(\theta - \hat{\theta}) \le \sigma^2 c\}$, where V is the $n \times p$ matrix with (i, j) element $\partial F(x_i; \theta)/\partial\theta_j$ evaluated at $\hat{\theta}$ (this follows by a straightforward computation of the expected information). For the purposes of the contours method, $f(\theta)$ is the density of an $N(\hat{\theta}, \sigma^2(V'V)^{-1})$ random variable and g is as in (2.2). The Fieller–Creasy problem can be cast as a nonlinear regression problem. Suppose that y_i is distributed as $N(\theta_1 x_i + \theta_1\theta_2(1 - x_i), \sigma^2)$, where $x_i = 1$ for population 1 and $x_i = 0$ for population 2. Then population 1 has unknown mean θ_1, population 2 has unknown mean $\theta_1\theta_2$, and the ratio of the population means is θ_2. Let $y_1, \ldots, y_n$ be observations on population 1, and let $y_{n+1}, \ldots, y_{2n}$ be observations on population 2.

The likelihood function is proportional to

$$\exp\{-n[(\theta_1 - \hat{\theta}_1)^2 + (\theta_1\theta_2 - \hat{\theta}_1\hat{\theta}_2)^2]/2\sigma^2\},$$

and

$$\hat{\theta}_1 = \sum_{i=1}^{n} y_i/n \quad \text{and} \quad \hat{\theta}_1\hat{\theta}_2 = \sum_{i=n+1}^{2n} y_i/n$$

give the MLE's for θ_1 and θ_2, which are sufficient statistics for this problem. For illustration purposes, suppose that $(\hat{\theta}_1, \hat{\theta}_2) = (3, 0)$ was observed and that $\sigma^2/n = 1$. Then for $\alpha = .05$ and $c = 5.99$, the upper 5% point of $\chi^2(2)$, the exact and approximate likelihood regions are $\{\theta \mid g(\theta) \le K_4\}$ and $\{\theta \mid f(\theta) \le K_4\}$, respectively, where $K_4 = 5.99 \exp(-5.99^2/12)/4\pi$. These contours are presented in Figure 1.

By construction, the intrinsic curvature of the solution locus $\eta(\theta) = ((\theta_1 x_i + \theta_1\theta_2)(1 - x_i))_{n\times 1}$ is identically zero.

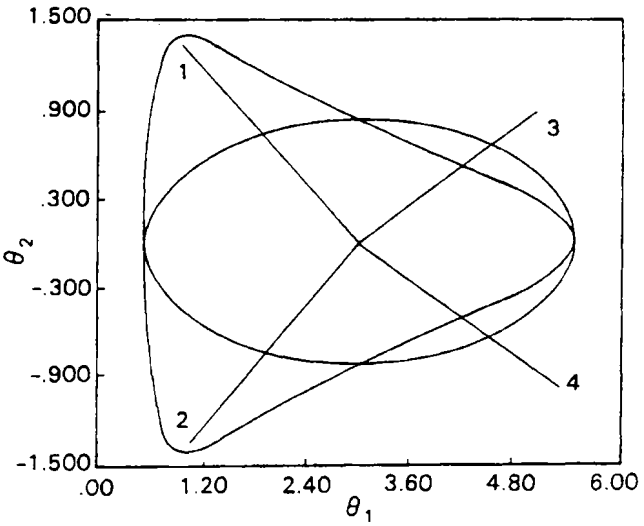

Figure 1. Contours of the Exact and Approximate Likelihoods for the Fieller–Creasy Problem. The wedge-shaped region is the exact region, and the ellipse is the approximate region. The digits on the plot label directions; for example, direction 3 points toward the northeast corner of the plot. [A slightly different version of this figure appeared in Cook and Witmer (1985).]

From Cook and Witmer's computation, the parameter–effects curvature is .33 and the guide value is .41. Bates and Watts's criterion indicates that the approximation shown in Figure 1 should be reasonable, yet apparently it is not. The contours method shows that points contained in the exact 95% likelihood region would be excluded by approximate regions with nominal confidence coefficients smaller than .99995 (directions 1 and 2 in Fig. 1), and points excluded from the exact 95% region would be included in approximate regions with nominal confidence levels larger than .8476 (directions 3 and 4 in Fig. 1).

Again, as in Section 3.1, we have the question of whether this approximation is a good approximation. I give the same answer as before. In this case, the approximate region resembles the exact region poorly enough that a 99.995% approximate region would be needed to contain the entire exact region. For some applications, a user might consider this adequate, but only the user can judge.

4. DISCUSSION

How do the two accuracy assessment methods in Section 2 measure up to the call for operational methods and to the other techniques? First I will discuss the other techniques, and then I will evaluate the two new methods. The most widely known measures for assessing the accuracy of an approximation are its rate of convergence and the Bates and Watts curvatures. [Two other approaches mentioned earlier, Jennings's (1986) and Minkin's (1983), are essentially precursors to the integrals and contours methods.]

The convergence rate of an approximation is a measure of the rate at which its error diminishes as the sample size increases. A typical statement is "$g(y) = f(y)\{1 + O(n^{-a})\}$," which means that there exists a constant k and a number $N(k)$ (both invariably unknown) such that $g(y)/f(y) - 1 < k/n^a$ for all $n > N(k)$. The idea is that approximations with errors $o(n^{-a})$, $O(n^{-a})$ or their stochastic counterparts are better for larger values of a because such approximations converge more rapidly to the exact function or statistic. This is true, but for a given model, data set, and approximation, a convergence rate does not provide relevant information about the size of the errors one risks in using the approximation to make the usual sort of statistical statement. For example, it is trivial to construct examples in which for a given model, sample size, and sample, the normal approximation to the posterior distribution—whose error converges at the rate $n^{-1/2}$ (Zellner 1971, p. 33)—is arbitrarily inaccurate. The point is that in any given situation, convergence rates cannot tell how well an approximation is working because they do not use information about the model, data, and prior at hand; in this sense, a convergence rate is not an operational assessment of accuracy.

Bates and Watts's curvature measures were discussed in Section 3.2. Unlike convergence rates, the curvatures do use information about the model and data at hand. But in the example in Section 3.2 the curvatures failed to signal the problematic approximation of the wedge-shaped region by the elliptical region. Cook and Witmer (1985) gave other examples for which excellent approximations are declared inadequate by the Bates–Watts measures. The difficulty with these measures is twofold. First, the likelihood is determined by the solution locus, so derivatives of the solution locus (i.e., the curvatures) convey information about the likelihood. But the derivatives of the solution locus are complicated functions of the derivatives of the likelihood, and if we are interested in the latter, the information conveyed by the former may not be in a desirable form. It is difficult to understand the information that the curvatures convey.

Second, the curvature methods are themselves defined by making an approximation, namely a quadratic Taylor series approximation to the solution locus. Although this does not mean that the curvatures are necessarily uninformative or misleading, it does mean, as noted in Section 1, that overcoming the shortcomings of the Taylor series approximation would require a redefinition of the method and not simply a refinement.

The integrals and contours methods presented in this article avoid these difficulties. Both methods use information not only about the problem at hand, but about the entity of interest, namely the exact density (for the integrals method) and the exact likelihood (for the contours method). The measures produced by the methods are interpretable: the integrals method provides information about the percent error of the approximation, and the contours method tells how large an approximate region would need to be to contain the exact region and how small it would need to be to be contained in the exact region. No approximations are used in defining the methods, although approximations are used in computing them.

[*Received July 1985. Revised April 1986.*]

REFERENCES

Bates, D. M., and Watts, D. G. (1980), "Relative Curvature Measures of Nonlinearity" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 42, 1–25.

Cook, R. D., and Witmer, J. A. (1985), "A Note on Parameter-Effects Curvature," *Journal of the American Statistical Association*, 80, 872–878.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Fisher, R. A. (1973), *Statistical Methods and Scientific Inference* (3rd ed.), New York: Hafner Press.

Gallant, A. R. (1975), "Testing a Subset of the Parameters of a Nonlinear Regression Model," *Journal of the American Statistical Association*, 70, 927–932.

Geisser, S. (1970), "Bayesian Analysis of Growth Curves," *Sankhya*, Ser. A, 32, 53–64.

Hodges, J. S. (1985a), "Assessing the Accuracy of Normal Approximations," Technical Report 449, University of Minnesota, School of Statistics.

——— (1985b), "Methods for Assessing the Accuracy of Some Standard Approximations," unpublished Ph.D. dissertation, University of Minnesota, School of Statistics.

Jennings, D. E. (1986), "Judging Inference Adequacy in Logistic Regression," *Journal of the American Statistical Association*, 81, 471–476.

Lee, J. C., and Geisser, S. (1972), "Growth Curve Prediction," *Sankhya*, Ser. A, 34, 393–412.

——— (1975), "Applications of Growth Curve Prediction," *Sankhya*, Ser. A, 37, 239–256.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, New York: Academic Press.

Minkin, S. (1983), "Assessing the Quadratic Approximation to the Log Likelihood Function in Non-normal Linear Models," *Biometrika*, 70, 367–372.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley.