Biometrika (1992), **79**, 2, *pp.* 231–45
Printed in Great Britain

More accurate confidence intervals in exponential families

BY THOMAS DICICCIO AND BRADLEY EFRON
Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

SUMMARY

Fisher's theory of maximum likelihood estimation routinely provides approximate confidence intervals for a parameter of interest θ, the standard intervals $\hat{\theta} \pm z_\alpha \hat{\sigma}$, where $\hat{\theta}$ is the maximum likelihood estimator, $\hat{\sigma}$ is an estimate of standard error based on differentiation of the log likelihood function, and z_α is a normal percentile point. Recent work has produced systems of better approximate confidence intervals, which look more like exact intervals when exact intervals exist, and in general have coverage probabilities an order of magnitude more accurate than the standard intervals. This paper develops an efficient and dependable algorithm for calculating highly accurate approximate intervals on a routine basis, for parameters θ defined in the framework of a multiparameter exponential family. The better intervals require only a few times as much computational effort as the standard intervals. A variety of numerical and theoretical arguments are used to show that the algorithm works well, and that the improvement over the standard intervals can be striking in realistic situations.

Some key words: Bias correction; Bootstrap BC_a limit; Edgeworth expansion; Empirical exponential family; Logistic regression; Nonparametric confidence limit; Second-order accuracy; Skewness correction.

1. INTRODUCTION

Fisher's papers of the 1920's on maximum likelihood estimation were a landmark for applied statistics as well as for statistical theory. In almost any parametric situation the maximum likelihood method automatically produces an efficient estimate $\hat{\theta}$ for a parameter of interest θ, as well as an estimate $\hat{\sigma}$ of standard error for $\hat{\theta}$. Together, $\hat{\theta}$ and $\hat{\sigma}$ give reasonable approximate confidence intervals for θ, the standard intervals

$$\theta \in \hat{\theta} \pm z_\alpha \hat{\sigma}, \tag{1·1}$$

z_α being the $(1-\alpha)$ quantile of the standard normal distribution. Their use in statistical applications vastly outweighs more exact methods, if for no other reason than the absence of exact intervals for most problems.

There has been considerable recent progress on better approximate confidence intervals. A variety of methods have been proposed that improve upon the standard intervals. These methods produce approximate intervals that look more like exact intervals when exact intervals exist, and that have better coverage accuracy than (1·1) in general. Some of the authors involved in this work are Welch & Peers (1963), Barndorff-Nielsen & Cox (1979), Withers (1983), McCullagh (1984), Barndorff-Nielsen (1986), Cox & Reid (1987), Efron (1987), Hall (1988), DiCiccio & Tibshirani (1987) and Konishi (1991).

So far all of this theoretical activity has had only modest impact on statistical applications, even though the potential benefits are impressive, as will be shown. Our main goal in this paper is to provide an efficient, dependable, and theoretically sound algorithm

for calculating highly accurate confidence intervals on a routine basis. The algorithm applies to real-valued parameters θ defined in the context of a multiparameter exponential family. This framework includes most of the familiar parametric situations: normal, binomial, Poisson, gamma, multinomial, Gaussian regression and analysis of variance models, logistic regression, contingency tables, log linear models, multivariate normal problems, Markov chains, etc. It will turn out to include also a class of nonparametric problems, as shown in § 4.

As a first example, Fig. 1 concerns data from $n = 26$ neurological patients, each of whom took two tests of spatial cognition, with the following results for scores (A, B): (42, 48), (33, 36), (16, 20), (39, 29), (38, 42), (36, 42), (15, 20), (33, 42), (20, 22), (43, 41), (34, 45), (22, 14), (7, 6), (15, 0), (34, 33), (29, 28), (41, 34), (13, 4), (38, 32), (25, 24), (27, 47), (41, 41), (28, 24), (14, 26), (28, 30), (40, 41). Two parameters are considered, the Pearson correlation coefficient between the two tests, and the variance of A. The horizontal axes of Fig. 1 indicate the intended coverage of the intervals, ranging from 0·68 to 0·95. All of the intervals assume that the 26 pairs of scores are independently sampled from a bivariate normal distribution. Lower and upper endpoints of the intervals are plotted in the vertical direction, moving further apart as the coverage increases. The dashed lines are the endpoints of the standard intervals (1·1), and as such they are symmetrically placed above and below the maximum likelihood estimate $\hat{\theta}$.

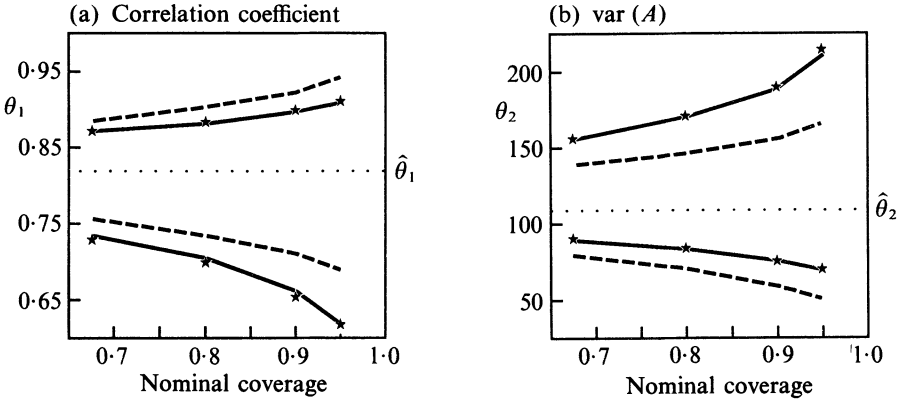

Fig. 1. Exact and approximate confidence intervals for parameters $\theta_1 = \text{corr}(A, B)$ and $\theta_2 = \text{var}(A)$, assuming a bivariate normal model for the spatial test data. Dashed curves are standard intervals (1·1); solid curves are ABC intervals, the approximate bootstrap confidence intervals constructed by the algorithm described in this paper. Stars indicate exact confidence intervals based on the bivariate normal model for coverage 0·68, 0·80, 0·90, 0·95; dotted line indicates maximum likelihood estimate.

The solid curves in Fig. 1 indicate the ABC intervals, produced by the algorithm discussed in this paper; ABC abbreviates 'approximate bootstrap confidence', or 'approximate BC_a'. However the ABC algorithm does not involve Monte Carlo simulations. Instead, the bootstrap calculations are approximated by analytic methods. The amount of calculation for the ABC curves is only a few times more than that for the standard intervals.

Exact confidence intervals are available in this situation. The stars in Fig. 1 denote the endpoints of the exact $1-2\alpha$ intervals, for $1-2\alpha = 0{\cdot}68, 0{\cdot}80, 0{\cdot}90, 0{\cdot}95$. Two facts are obvious: the standard intervals err substantially for both problems, while the ABC intervals are close to being exactly correct. Section 5 shows that the ABC intervals are second-order accurate; they have asymptotic error probability $\alpha + O(n^{-1})$ in each tail, compared to

the intended value α. The standard intervals, which are only first-order accurate, have error probability $\alpha + O(n^{-\frac{1}{2}})$. The difference between first- and second-order accuracy can be quite striking in small samples, as Fig. 1 demonstrates.

Bivariate normality is a questionable assumption for the spatial test data. This does not affect the comparisons in Fig. 1. All of the intervals are functions of the bivariate normal sufficient statistics, two sample means, two variances, and a covariance, given which the intervals are as shown, irrespective of how normal or nonnormal the 26 data points appear. A bad model can make all of the intervals wrong, of course. Figure 3 displays nonparametric intervals for the correlation coefficient and var (A), based on the empirical exponential families discussed in § 4. Switching to a nonparametric model has little effect on the correlation coefficient intervals, but a great effect on the inferences for var (A).

The goal of improving on the standard intervals can be approached in many different ways. The theoretical work of the previously mentioned authors follows four main lines, Edgeworth expansion, saddlepoint methods, corrected likelihoods, and bootstrap computations. T. DiCiccio and B. Efron, in an unpublished longer version of this paper, show that all of these approaches agree with each other, and with the ABC intervals, at the level of second-order accuracy. In this sense, the line of approach to second-order accuracy does not matter.

In terms of accuracy, dependability, and efficiency in small samples, the approach matters a great deal. The ABC algorithm has performed consistently well in a wide variety of numerical tests. It is numerically efficient, and requires only minor adjustments for different applications within the exponential family framework. A program for the ABC algorithm written in the language S of Becker, Chambers & Wilks (1988) is available from the authors.

We believe that more ambitious approaches to approximate confidence intervals could also be algorithmized, applying to more general parametric families and giving third-order accuracy. So far though, this more ambitious plan has proved difficult to execute in any general sense. As a practical point, second-order accuracy is often good enough to give the kind of results seen in Fig. 1, leaving little room for third-order improvements.

2. ABC INTERVALS

The ABC intervals are approximations to the bootstrap BC_a intervals of Efron (1987), using analytic methods to avoid the BC_a's Monte Carlo calculations. The ABC algorithm gives second-order accurate confidence intervals for real-valued parameters defined in multiparameter exponential families. This section describes the algorithm, with emphasis on the three corrections it makes to the standard intervals (1·1). We begin here with a brief review of exponential families; see also Efron (1978).

Suppose that a data vector y has a family of possible density functions indexed by the parameter vector μ. Let $\theta = t(\mu)$ indicate a real-valued parameter of particular interest, estimated by its maximum likelihood estimator $\hat{\theta} = t(\hat{\mu})$, where $\hat{\mu}$ is the maximum likelihood estimator of the entire vector μ. The function $t(\mu)$ must be smoothly defined, having all of its second derivatives. We will assume that the parametric family of densities for y is an exponential family, with p-dimensional natural parameter vector η, so the density can be expressed as

$$g_\eta(y) = e^{\eta' y - \psi(\eta)}. \tag{2·1}$$

The p-dimensional data vector y is usually a sufficient statistic obtained from a larger data set involving n components, as shown in the specific examples of § 3.

The expectation parameter vector $\mu \equiv E_\eta(y)$ can be obtained by differentiating $\psi(\eta)$,

$$\mu = \dot{\psi}(\eta) \equiv \{\partial\psi(\eta)/\partial\eta_i\}'. \tag{2·2}$$

The vector μ indexes the exponential family just as well as does η, since the mapping from η to μ is one-to-one. We will use notation like $g_\mu(y)$ and $E_\mu(y)$ when convenient, meaning $g_\eta(y)$ and $E_\eta(y)$ for the value of η corresponding to μ.

The $p \times p$ covariance matrix of y, $\dot{\Sigma}(\mu)$, is given by the second derivative matrix of ψ,

$$\dot{\Sigma}(\mu) = \ddot{\psi}(\eta) \equiv \left(\frac{\partial^2\psi}{\partial\eta_i\partial\eta_j}\right) \quad (i,j = 1,\ldots,p). \tag{2·3}$$

Likewise, third derivatives of $\psi(\eta)$ give third-order cumulants, etc.

The maximum likelihood estimator of the expectation vector μ is y, say $\hat{\mu} = y$, so the maximum likelihood estimator of a real-valued parameter $\theta = t(\mu)$ is $\hat{\theta} = t(\hat{\mu}) = t(y)$. The familiar 'delta method' approximation for the standard error of $\hat{\theta}$ is

$$\sigma = \{\dot{t}(\mu)'\dot{\Sigma}(\mu)\dot{t}(\mu)\}^{\frac{1}{2}}, \tag{2·4}$$

where $\dot{t}(\mu)$ is the gradient vector $(\partial t/\partial\mu_i)'$, this also being the Cramér–Rao lower bound for the standard deviation of an unbiased estimator of θ. In practice, $\hat{\mu} = y$ substitutes for μ in (2·4), giving an estimated standard error for $\hat{\theta}$

$$\hat{\sigma} = \{\dot{t}(\hat{\mu})'\hat{\dot{\Sigma}}\dot{t}(\hat{\mu})\}^{\frac{1}{2}}, \tag{2·5}$$

$\hat{\dot{\Sigma}} \equiv \dot{\Sigma}(\hat{\mu}) = \dot{\Sigma}(y)$. Because $\hat{\dot{\Sigma}}$ is the inverse of the estimated Fisher information matrix for $\hat{\mu}$, expression (2·5) is the usual estimate of standard error for $\hat{\theta}$ based on differentiating the log likelihood function.

The standard intervals (1·1) depend on the calculation of the two quantities $(\hat{\theta}, \hat{\sigma})$. The ABC intervals require three more quantities, called (a, z_0, c_q), whose calculation and interpretation are discussed below. We will present two forms of the ABC confidence limits, one easier to calculate and the other enjoying a theoretical advantage. Let $\hat{\theta}[1-\alpha]$ indicate the endpoint of an approximate $(1-\alpha)$ one-sided upper confidence interval for θ, so that $(\hat{\theta}[\alpha], \hat{\theta}[1-\alpha])$ is an approximate $(1-2\alpha)$ two-sided interval. Subscript notations such as $\hat{\theta}_{\mathrm{ABC}}[1-\alpha]$ will indicate different methods.

Definition 1. The quadratic ABC confidence limit for θ, denoted by $\hat{\theta}_{\mathrm{ABC}_q}[1-\alpha]$, is constructed from $(\hat{\theta}, \hat{\sigma}, a, z_0, c_q)$ and $z_\alpha = \Phi^{-1}(1-\alpha)$ as follows:

$$w \equiv z_0 + z_\alpha, \quad \lambda \equiv \frac{w}{(1-aw)^2}, \quad \xi \equiv \lambda + c_q\lambda^2, \quad \hat{\theta}_{\mathrm{ABC}_q}[1-\alpha] = \hat{\theta} + \hat{\sigma}\xi. \tag{2·6}$$

Definition 2. The ABC confidence limit for θ, denoted by $\hat{\theta}_{\mathrm{ABC}}[1-\alpha]$, is constructed as follows:

$$w \equiv z_0 + z_\alpha, \quad \lambda \equiv \frac{w}{(1-aw)^2}, \quad \hat{\delta} \equiv \hat{\dot{\Sigma}}\dot{t}(\hat{\mu}), \quad \hat{\theta}_{\mathrm{ABC}}[1-\alpha] = t(\hat{\mu} + \lambda\hat{\delta}/\hat{\sigma}). \tag{2·7}$$

We will refer to both definitions as ABC confidence intervals, using $\hat{\theta}_{\mathrm{ABC}}$ or $\hat{\theta}_{\mathrm{ABC}_q}$ to distinguish between them when necessary.

The asymptotic difference between $\hat{\theta}_{\mathrm{ABC}}$ and $\hat{\theta}_{\mathrm{ABC}_q}$ is negligible but in our examples $\hat{\theta}_{\mathrm{ABC}}$ seems to perform a little better, for reasons suggested below. The ABC curves in Fig. 1 were obtained from (2·7), though in this case there was little difference between $\hat{\theta}_{\mathrm{ABC}}$ and $\hat{\theta}_{\mathrm{ABC}_q}$.

Formula (2·7) enjoys a nice invariance property. Suppose we change our parameter of interest from $\theta = t(\mu)$ to some smooth monotonically increasing function of θ, say $\phi = m(\theta)$. Then it is possible to show that the $\hat{\theta}_{\text{ABC}}$ limits transform correctly,

$$\hat{\phi}_{\text{ABC}}[1-\alpha] = m(\hat{\theta}_{\text{ABC}}[1-\alpha]). \tag{2·8}$$

This is not true for the $\hat{\theta}_{\text{ABC}_q}$ limits, a disadvantage which can sometimes limit their accuracy.

The disadvantage of (2·7) is computational. The recomputations of $t(.)$ for $\hat{\theta}_{\text{ABC}_q}$ are local, in the sense that they only require evaluating $t(\mu)$ for μ near $\hat{\mu} = y$. This simplifies the numerical work by avoiding boundary problems and any form of global searching or maximization. The $\hat{\theta}_{\text{ABC}}$ endpoints (2·7) require two additional nonlocal recomputations of $t(.)$. This can be annoying, as in the logistic regression example of § 3, though it is conceptually simple. Saddlepoint-based methods for confidence intervals, like those of Barndorff-Nielsen (1986), are strongly nonlocal, requiring repeated global searches and maximizations. This is an impediment to their routine use, though an impediment we can expect to see overcome with further development.

If $(a, z_0, c_q) = (0, 0, 0)$, then $\hat{\theta}_{\text{ABC}_q}[1-\alpha]$ in (2·6) reduces to the standard interval endpoint $\hat{\theta}_{\text{stan}}[1-\alpha] = \hat{\theta} + \hat{\sigma} z_\alpha$. The same is true of $\hat{\theta}_{\text{ABC}}[1-\alpha]$ if $(a, z_0) = (0, 0)$ and $t(\hat{\mu} + \lambda \hat{\delta}/\hat{\sigma})$ is linear in λ. Nonzero values of (a, z_0, c_q) produce second-order corrections to the standard intervals. In Fig. 1, where the corrections are sizeable, (a, z_0, c_q) equals $(0, -0{\cdot}080, -0{\cdot}161)$ for the correlation coefficient and $(0{\cdot}092, 0{\cdot}243, 0)$ for var (A).

Each of the quantities (a, z_0, c_q) makes a different correction to the standard intervals: z_0 corrects for bias in the maximum likelihood estimator; c_q corrects for nonlinearity in the function $\theta = t(\mu)$; the acceleration a corrects for the fact that the standard error of $\hat{\theta}$ is nonconstant, varying with μ. Asymptotically $(a, z_0, c_q) \to (0, 0, 0)$, at rate $n^{-\frac{1}{2}}$ in repeated sampling situations, but in finite samples the corrections can be quite important, as seen in Fig. 1.

Considerably more information about (a, z_0, c_q) is given by Efron (1985, 1987) and in the authors' unpublished technical report. Two important points are as follows: the ABC endpoints are analytical approximations to the full bootstrap confidence intervals called BC_a of Efron (1987), approximations which are possible because we are dealing here with smooth distributions and parameters; these approximations make heavy use of Stein's (1956) least favourable family construction. For instance $\hat{\delta} = \hat{\dot{\ell}} \hat{i}(\hat{\mu})$ in (2·7) is the least favourable direction in the expectation space of family (2·1).

The ABC algorithm computes (a, z_0, c_q) in terms of numerical second derivatives. The acceleration constant a in (2·6) and (2·7) is calculated from the formula

$$a = \frac{1}{6\hat{\sigma}^3} \left[\frac{\partial^2}{\partial h^2} \dot{\ell}' \mu(\hat{\eta} + h\dot{\ell}) \right]_{h=0}, \tag{2·9}$$

where $\dot{\ell} \equiv \dot{\ell}(\hat{\mu})$, $\hat{\eta}$ is the maximum likelihood estimator of the natural parameter vector η, and $\mu(\hat{\eta} + h\dot{\ell})$ is the expectation vector corresponding to natural parameter $\hat{\eta} + h\dot{\ell}$. This is the same expression for a as formula (6·7) of Efron (1987).

The next second derivatives calculated by the ABC algorithm relate to the bias $E(\hat{\theta}) - \theta$. A quadratic Taylor series expansion of $\theta = t(\mu)$ around $\mu = \hat{\mu} = y$ gives approximate bias

$$\hat{b} = \text{tr}\,(\tfrac{1}{2} \ddot{t} \hat{\dot{\ell}}), \tag{2·10}$$

where $\ddot{t}$ is the $p \times p$ matrix with elements $\ddot{t}_{ij} = \partial^2 t(\mu)/\partial \mu_i \partial \mu_j|_{\mu=\hat{\mu}}$. The eigen decomposition $\hat{\dot{\ell}} = \Gamma D \Gamma'$, with $D = \text{diag}\,(d_1, \ldots, d_p)$ and Γ equal to an orthogonal matrix $(\gamma^1, \ldots, \gamma^p)$,

yields the formula

$$\hat{b} = \tfrac{1}{2} \sum_{j=1}^{p} \left[\frac{\partial^2}{\partial h^2} t(\hat{\mu} + h d_j^{\frac{1}{2}} \gamma^j) \right]_{h=0}. \tag{2·11}$$

The final second derivative involved in the ABC calculations gives the quadratic coefficient c_q,

$$c_q = \frac{1}{2\hat{\sigma}} \left[\frac{\partial^2}{\partial h^2} t\left(\hat{\mu} + \frac{\hat{\delta}}{\hat{\sigma}} h \right) \right]_{h=0}. \tag{2·12}$$

This coefficient measures the nonlinearity of the function $\theta = t(\mu)$ as we move in the least favourable direction. Let $\theta(\lambda) \equiv t(\hat{\mu} + \lambda \hat{\delta}/\hat{\sigma})$. A quadratic Taylor series expansion gives

$$\theta(\lambda) \doteq \hat{\theta} + \hat{\sigma}(h + c_q h^2); \tag{2·13}$$

c_q measures the ratio of the quadratic term to the linear term in $\{\theta(h) - \hat{\theta}\}/\hat{\sigma}$. The size of c_q does not affect the standard intervals, which treat every function $t(\mu)$ as if it were linear, but it has an important effect on more accurate confidence intervals. Note that approximation (2·13) relates $\hat{\theta}_{\mathrm{ABC}}$ to $\hat{\theta}_{\mathrm{ABC}_q}$ in (2·6) and (2·7).

The bias correction constant z_0 is a function of a, $\hat{b}$ and c_q. Define

$$\hat{\gamma} = \hat{b}/\hat{\sigma} - c_q. \tag{2·14}$$

Then

$$z_0 \doteq \Phi^{-1}\{2 \cdot \Phi(a) \cdot \Phi(-\hat{\gamma})\} \doteq a - \hat{\gamma}. \tag{2·15}$$

Formula (2·15) is a straightforward generalization of (8.8) of Efron (1987), with $a = z_{01}$ and $\hat{\gamma} = -z_{02}$ in the notation there: $\hat{\gamma}$ is the total curvature of the level surface $\{\mu : t(\mu) = \hat{\theta}\}$ in the expectation space, evaluated at $\hat{\mu} = y$ in the metric $\hat{\dot{\Sigma}}^{-1}$; it is the quantity $\mathrm{tr}\,(d_\eta)$ in Theorem 1 of Efron (1985). The greater the curvature, the more biased is $\hat{\theta}$. Either form of (2·15) approximates the original definition of z_0 of Efron (1987) sufficiently well to preserve the second-order accuracy of the BC_a formulae. The definition of z_0 is more like a median bias than a mean bias, which is why z_0 involves quantities other than $\hat{b}$.

How much computation is required for the ABC intervals? The algorithm begins by numerically evaluating $\dot{t} = \dot{t}(\hat{\mu})$. This requires $2p$ recomputations of $t(.)$, two for each of the first derivatives

$$\left[\frac{\partial t(\mu)}{\partial \mu_i} \right]_{\mu = \hat{\mu}} \doteq \frac{t(\hat{\mu} + h e_i) - t(\hat{\mu} - h e_i)}{2h},$$

e_i being the ith coordinate vector. The vector $\dot{t}$ gives $\hat{\sigma} = \{\dot{t}' \hat{\dot{\Sigma}} \dot{t}\}^{\frac{1}{2}}$. Then the $p+2$ second derivatives in (2·9), (2·11) and (2·12) are calculated, each requiring two recomputations of $t(.)$. Altogether $4p+4$ recomputations of $t(.)$ are required to compute the quadratic ABC limits (2·6), compared with the $2p$ recomputations necessary for numerically evaluating the standard interval $\hat{\theta} \pm z_\alpha \hat{\sigma}$. In complicated situations the recomputations of $t(.)$ dominate calculational expense, so it is fair to say that the ABC_q limits require less than three times as much numerical effort as the standard limits.

Despite the large number of approximations involved in their derivation, the ABC intervals closely matched the full bootstrap BC_a intervals in all the cases considered. For the spatial test data correlation coefficient, the BC_a interval with $\alpha = 0{\cdot}05$, based on $N_B = 4800$ bootstrap replications, was $(0{\cdot}670 \pm 0{\cdot}004,\ 0{\cdot}903 \pm 0{\cdot}001)$, compared with the

ABC interval (0·668, 0·901). The ± values indicate the Monte Carlo standard errors based on $N_B = 4800$ replications. The corresponding intervals for the parameter var (A) were $(76{\cdot}4 \pm 0{\cdot}5,\ 198{\cdot}5 \pm 2{\cdot}6)$ compared with (76·1, 193·5).

The advantage gained by restricting attention to exponential families is more computational than theoretical: the number of numerical derivatives required is small; only first and second derivatives are needed, avoiding the more demanding task of evaluating numerical third derivatives; exact expressions are available for the moments of y, which reduces the number of approximations made; and the only function, besides $t(y)$, needed to implement the algorithm is $\mu(\eta)$, which has an easy-to-program form in all of the familiar exponential families. The ABC algorithm gives a preferred place to the expectation space μ, as seen in (2·7). An attempt to shift the emphasis to the η space provided noticeably less accurate numerical results, even though the intervals were just as good in theory.

3. Familiar exponential families

This section explicitly describes the calculation of ABC confidence intervals for parameters in the most widely used exponential families: Poisson, multinomial, binomial, logistic regression, gamma and multivariate normal. The details vary from family to family because of differences in how the families are traditionally written.

Five items must be specified in order to calculate the ABC intervals: the function $t(.)$ that describes the parameter of interest $\theta = t(\mu)$; the sufficient vector $y = \hat{\mu}$, (2·1); the maximum likelihood estimator of the covariance matrix, $\hat{\dot{\Sigma}} = \dot{\Sigma}(\hat{\mu})$; the maximum likelihood estimator of the natural parameter, $\hat{\eta}$; and the function $\mu(\eta)$ mapping the natural parameter vector η to the expectation vector μ. We will describe y, $\hat{\dot{\Sigma}}$, $\hat{\eta}$ and $\mu(\eta)$ for the families listed above. The function $t(.)$ depends of course on the particular application.

Poisson. The observed data consists of p independent Poisson variates, each with its own unknown mean,

$$y_j \sim \text{Po}(\mu_j) \quad (j = 1, \ldots, p). \tag{3·1}$$

In this case the data vector $y = (y_1, \ldots, y_p)'$ has mean vector $\mu = (\mu_1, \ldots, \mu_p)'$, and covariance matrix diag (μ) = the $p \times p$ diagonal matrix with jth diagonal element μ_j, so $\hat{\dot{\Sigma}} = \dot{\Sigma}(\hat{\mu}) = \text{diag}(y)$. The natural parameter vector η has components $\eta_j = \log(\mu_j)$, so $\hat{\eta}_j = \log(y_j)$, and the function $\mu(\eta)$ is $\mu_j = e^{\eta_j}$, for $j = 1, \ldots, p$. In order to avoid boundary problems, in the multinomial and binomial cases as well as here, we need all $y_j \neq 0$. Zero observations can usually be handled by the familiar tactic of substituting a small positive value, like $\frac{1}{2}$, but the resulting intervals should be checked for sensitivity to the substituted value.

Multinomial. Suppose that n subjects are independently classified into one of p disjoint categories, with unknown probability π_j for category $j = 1, \ldots, p$. The sufficient vector is $y = (y_1, \ldots, y_p)'$, y_j being the number of subjects classified into category j; y has the multinomial distribution on p categories, n draws, true probability vector π, $y \sim \text{Mult}_p(n, \pi)$. The mean vector μ equals $n\pi$, so μ_j/n is the true probability π_j for category j.

Straightforward calculations show that the ABC intervals for $\theta = t(\mu)$ are exactly the same for $y \sim \text{Mult}_p(n, \mu/n)$ as for y having the Poisson distribution (3·1). We can use the Poisson specifications to handle the multinomial case, with one point of caution: the

function $t(.)$ must be homogeneous, i.e. satisfy $t(cy) = t(y)$ for $c > 0$. If this is not true, then the homogenized form $t_h(y)$ should be substituted for $t(y)$,

$$t_h(y) \equiv t\left(ny \Big/ \sum_{j=1}^{p} y_j \right).$$

If $t(y) = y_1/n$ for instance, then $t_h(y) = y_1/\Sigma y_j$, where the sum is over the range $j = 1, \ldots, p$.

Binomial. We observe independent binomial variates, $y_j \sim \text{Bi}(n_j, \pi_j)$ $(j = 1, \ldots, p)$. The data vector $y = (y_1, \ldots, y_p)'$ has mean vector $\mu = (\mu_1, \ldots, \mu_p)'$ with $\mu_j = n_j\pi_j$. The covariance matrix is

$$\dot{\Sigma}(\mu) = \text{diag}\{n_j\pi_j(1-\pi_j)\} = \text{diag}\{\mu_j(1-\mu_j/n_j)\},$$

so $\hat{\dot{\Sigma}} = \text{diag}\{y_j(1-y_j/n)\}$. The natural parameter vector η has components

$$\eta_j = \log\{\pi_j/(1-\pi_j)\} = \log\{\mu_j/(n_j-\mu_j)\},$$

so $\hat{\eta}_j = \log\{y_j/(n_j-y_j)\}$; the mapping $\mu(\eta)$ from η to μ is $\mu_j = n_j/(1+e^{-\eta_j})$ for $j = 1, \ldots, p$.

Logistic regression. A standard logistic regression analysis links independent binomial observations by a linear regression model for the logistic parameters (Cox & Snell, 1989): we observe independent binomials

$$z_i \sim \text{Bi}(n_i, \pi_i) \quad (i = 1, \ldots, N), \tag{3·2}$$

and also $p \times 1$ covariate vectors x_i, which are the columns of the $p \times N$ matrix X. The logit for the ith case is assumed to be a linear function of x_i and an unknown regression vector η,

$$\log\{\pi_i/(1-\pi_i)\} = x_i'\eta \quad (i = 1, \ldots, N). \tag{3·3}$$

The model in (3·2) and (3·3) is a p-parameter exponential family with sufficient vector $y = Xz = \Sigma x_i z_i$, where the sum is over the range $i = 1, \ldots, N$, and natural parameter vector η. A standard logistic regression program provides the maximum likelihood estimator $\hat{\eta}$ of η. There is no closed-form expression for $\hat{\eta}$, but see (3·4) below. The estimated covariance matrix $\hat{\dot{\Sigma}}$ equals

$$X \text{ diag}\{n_i\hat{\pi}_i(1-\hat{\pi}_i)\}X' = \sum x_i n_i \hat{\pi}_i(1-\hat{\pi}_i)x_i',$$

where $\hat{\pi}_i$ is the maximum likelihood estimator for the ith probability,

$$\hat{\pi}_i = 1/(1+e^{-x_i'\hat{\eta}}).$$

Most logistic regression programs provide the matrix $\hat{\dot{\Sigma}}$ in their outputs. The function $\mu(\eta)$ is

$$\mu = \sum_{i=1}^{N} x_i n_i/(1+e^{-x_i'\eta}).$$

Parameters of interest in a logistic regression are often specified as functions $s(\eta)$ of the natural parameter η rather than the expectation vector μ. For instance we might have $\theta = \{1+\exp(-x_0'\eta)\}^{-1}$, the probability of a positive response at covariate vector x_0. For computation purposes in our algorithm we need to evaluate the implied function $t(y+\Delta) = s\{\hat{\eta}(y+\Delta)\}$ for various choices of the incremental vector Δ, where $\hat{\eta}(y+\Delta)$ is the maximum likelihood estimator of η corresponding to sufficient statistic $y+\Delta$. Since Δ is small, it is convenient to evaluate $s\{\hat{\eta}(y+\Delta)\}$ by Newton–Raphson updating, going from a preliminary guess $\hat{\eta}^{(1)}$ to an updated guess $\hat{\eta}^{(2)}$ according to

$$\hat{\eta}^{(2)} = \hat{\eta}^{(1)} + \hat{\dot{\Sigma}}^{-1}\{y+\Delta-\mu(\hat{\eta}^{(1)})\}. \tag{3·4}$$

A simpler and faster way to proceed, discussed in the authors' unpublished technical report, is to work directly with parameters specified as $\theta = s(\eta)$. Then (a, z_0, c_q) can be calculated from derivatives in the η space.

Gamma. Suppose we observe independent scaled gamma variates, each with its own unknown scale parameter, $y_i \sim \mu_i G_{n_i}/n_i$ $(i = 1, \ldots, p)$. Each y_i might be an average $\bar{x}_i = \Sigma x_{ij}/m_i$, the sum being over the range $j = 1, \ldots, m_i$, where $x_{ij} \sim \mu_i G_{\nu_i}/\nu_i$ with ν_i a known shape parameter, in which case $n_i = m_i\nu_i$. Then the ABC algorithm uses $y = (y_1, \ldots, y_p)'$, $S = \text{diag}\,(y_i^2/n_i)$, $\hat{\eta}_j = -n_j/y_j$ for $j = 1, \ldots, p$, and $\mu(\eta)_j = -n_j/\eta_j$ for $j = 1, \ldots, p$.

Multivariate normal. Suppose that the data consists of n independent observations from a d-dimensional multivariate normal distribution with unknown mean vector λ and unknown covariance matrix Γ, $x_i \sim N_d(\lambda, \Gamma)$ $(i = 1, \ldots, n)$, so $d = 1$ is the univariate normal case. The sufficient vector y has $p = d(d+3)/2$ coordinates, say $y' = (y_1', y_2')$ where y_1 is the d-dimensional vector of means and y_2 is the $d(d+1)/2$-dimensional vector of mean squares,

$$y_1 = \sum_{i=1}^{n} x_i/n = \bar{x}, \quad y_2 = \left(\sum_{i=1}^{n} x_i x_i'/n\right)^{(v)}.$$

Here $c = C^{(v)}$ indicates the $d(d+1)/2$-dimensional vector c obtained by ordering the super-diagonal elements of the symmetric $d \times d$ matrix C in any convenient way, say

$$c' = (C_{11}, C_{12}, C_{13}, \ldots, C_{22}, C_{23}, \ldots, C_{dd}).$$

The inverse mapping from c back to C will be denoted by $C = c^{(m)}$.

The mean value vector μ of y is similarly partitioned into $\mu' = (\mu_1', \mu_2')$, with

$$\mu_1 = \lambda, \quad \mu_2 = (\lambda\lambda' + \Gamma)^{(v)}. \tag{3·5}$$

The natural parameter vector $\eta' = (\eta_1', \eta_2')$ is given by

$$\eta_1 = n\Gamma^{-1}\lambda, \quad \eta_2 = n\,\{\text{diag}\,(\Gamma^{-1})/2 - \Gamma^{-1}\}^{(v)}, \tag{3·6}$$

where diag (A) indicates the diagonal matrix obtained from a square matrix A by setting its off-diagonal elements equal to zero. The mapping $\mu(\eta)$ from η to μ is expressed in two steps, first from η to (λ, Γ),

$$\Gamma = -n\,\{\text{diag}\,(\eta_2^{(m)}) + \eta_2^{(m)}\}, \quad \lambda = \frac{1}{n}\Gamma(\eta_1), \tag{3·7}$$

and then from (λ, Γ) to μ by (3·5). Likewise the mapping for μ to η, or from $\hat{\mu}$ to $\hat{\eta}$, is given by $\lambda = \mu_1$, $\Gamma^{-1} = (\mu_2^{(m)} - \mu_1\mu_1')^{-1}$, followed by (3·6).

In this case, and in others involving the more elaborate exponential families, it is easiest to calculate the covariance matrix $\hat{\dot{\Sigma}} = \dot{\Sigma}(\hat{\mu})$ needed in the ABC algorithm by numerical differentiation of $\mu(\eta)$, using

$$\hat{\dot{\Sigma}}_{p\times p} = \dot{\mu}(\hat{\eta}) = [\partial\mu_i/\partial\eta_j]_{\eta=\hat{\eta}}.$$

Alternatively, $\hat{\dot{\Sigma}}$ for the multivariate normal case can be calculated from lengthy but straightforward formulae.

As an example, we consider the score data from Mardia, Kent & Bibby (1979, pp. 3–4). The data consist of $d = 5$ test scores, for each of $n = 88$ students. Starting with the multivariate normal model, we consider two parameters of interest, θ_1 the maximum

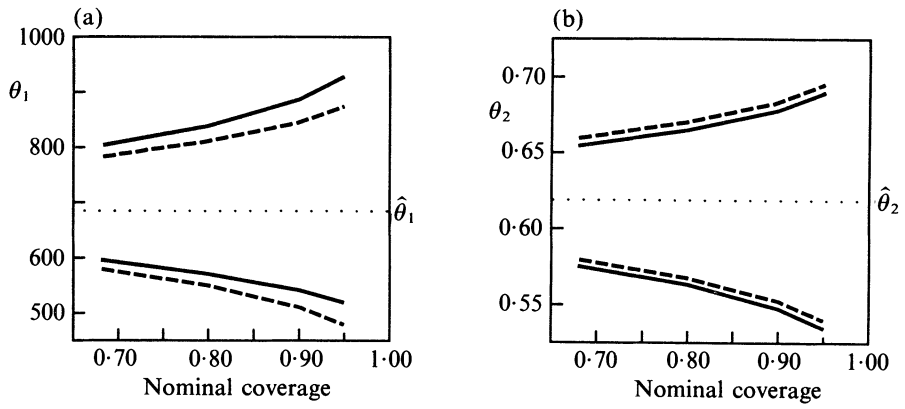

Fig. 2. Approximate confidence intervals for parameters θ_1 and θ_2 in the multivariate normal model $x_i \sim N_d(\lambda, \Gamma)$ $(i = 1, \ldots, n)$. (a) θ_1 = maximum eigenvalue of Γ; (b) $\theta_2 = \theta_1/\text{tr}\,(\Gamma)$. Data from Mardia, Kent & Bibby (1979, pp. 3–4), with $n = 88$ and $d = 5$. Dashed curves, standard intervals (1·1); solid curves, ABC intervals.

eigenvalue of Γ and $\theta_2 = \theta_1/\text{tr}\,(\Gamma)$, the proportion of the total variance attributable to the first principal direction. Figure 2 compares the ABC intervals with the standard intervals. For θ_1 the ABC intervals are shifted noticeably up from the standard, while θ_2 shows a smaller downward shift. In both cases, but particularly for θ_2, even the standard intervals are not usually available to the statistician.

4. NONPARAMETRIC CONFIDENCE INTERVALS

The ABC algorithm can be used to find second-order accurate nonparametric confidence intervals. Suppose that we are in the one-sample nonparametric situation, where a completely unknown probability distribution F has yielded a random sample of size n, $x = (x_1, \ldots, x_n)$. Corresponding to each x_j is a p-dimensional vector of pertinent statistics $v_j = V(x_j)$. The parameter of interest is some function of the expected value μ of v, $\theta = t(\mu)$, which we estimate by $\hat{\theta} = t(\bar{v})$, where $\bar{v} = \Sigma\, v_j/n$, and the sum is over the range $j = 1, \ldots, n$. For example, the x_j could be bivariate as in Fig. 1, $x_j = (A_j, B_j)$, with

$$v = V(x) = (A, B, A^2, AB, B^2)', \quad \theta = t(\mu) = \frac{\mu_4 - \mu_1\mu_2}{\{(\mu_3 - \mu_1^2)(\mu_5 - \mu_2^2)\}^{\frac{1}{2}}} = \text{corr}_F(A, B).$$

The BC_a algorithm gives a nonparametric second-order accurate confidence interval for θ based on $\hat{\theta}$ (Efron, 1987; Hall, 1988). We can approximate the BC_a interval with the ABC algorithms, avoiding all of the Monte Carlo calculations, by using a specially constructed exponential family called the empirical exponential family.

Let $y = \bar{v}$, and define $g_\eta(y) = e^{\eta' y - \psi(\eta)}$, (2·1), to be a density function with respect to the distribution $\hat{F}^n$, the product distribution of n independent copies of the empirical distribution $\hat{F}$ that puts probability $1/n$ on x_j $(j = 1, \ldots, n)$. In other words, the parameter vector η corresponds to a distribution F_η putting probability $g_\eta(y^*)/n^n$ on each of the n^n data sets $x^* = (x_1^*, \ldots, x_n^*)$, where each x_i^* is one of the original x_j points, and $y^* = \bar{v}^* = \Sigma_j V(x_j^*)/n$,

$$dF_\eta(x^*) = e^{\eta' y^* - \psi(\eta)} d\hat{F}^n(x^*). \tag{4·1}$$

This is the empirical exponential family. The stars distinguish the hypothetical random variable x^* from the actual data set x that gave $\hat{F}$ and y.

We can describe (4·1) as the family passing through the best-fitting nonparametric distribution $\hat{F}^n$, in the p directions determined by the vector of pertinent statistics v. Almost the same description applies to the parametric families we have discussed previously, except that they pass through the best-fitting parametric distribution $F^n_{\hat{\eta}}$ instead of $\hat{F}^n$. The maximum likelihood estimator of μ is $\hat{\mu} = y = \bar{v}$ in both the parametric and nonparametric families, so $\hat{\theta} = t(\bar{v})$ in both situations.

It is easy to calculate that

$$\psi(\eta) = n \log\left(\sum_{k=1}^{n} e^{\eta' v_k/n}/n\right)$$

in (4·1), so the mean $\mu(\eta) = \dot{\psi}(\eta)$ in (2·2) is

$$\mu(\eta) = \sum_{k=1}^{n} v_k e^{\eta' v_k/n} \Big/ \sum_{k=1}^{n} e^{\eta' v_k/n}. \tag{4·2}$$

Notice that $\mu(0) = \bar{v} = y$, so the maximum likelihood estimator of η is $\hat{\eta} = 0$. The estimated covariance matrix at $\hat{\eta} = 0$, from (2·3), is

$$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{n} (v_k - \bar{v})(v_k - \bar{v})'/n^2. \tag{4·3}$$

With these specifications, we can use the ABC algorithm to approximate the BC_a confidence interval for $\theta = t(\mu)$ based on $\hat{\theta} = t(\bar{v})$.

Figure 3 is a nonparametric version of Fig. 1, using empirical exponential families instead of the bivariate normal model. The results are nearly the same as before for the correlation coefficient, but not for var (A). Both the ABC and standard intervals are much narrower for var (A). This happens because the empirical distribution of the A scores looks more uniform than normal, suggesting that the normal model exaggerates the variability of the sample variance $\hat{\theta}$. In particular the estimated standard error $\hat{\sigma}$, (2·5), equals 30·34 for the normal model but only 20·99 nonparametrically. The nonparametric estimate of $\hat{\sigma}$ is $\{\dot{t}(\bar{v})'\hat{\mathbf{\Sigma}}\dot{t}(\bar{v})\}^{\frac{1}{2}}$, from (4·3).

The stars in Fig. 3 indicate the endpoints of the nonparametric BC_a intervals, as described in § 7 of Efron (1987), based on $N_B = 4800$ bootstrap replications. We see that

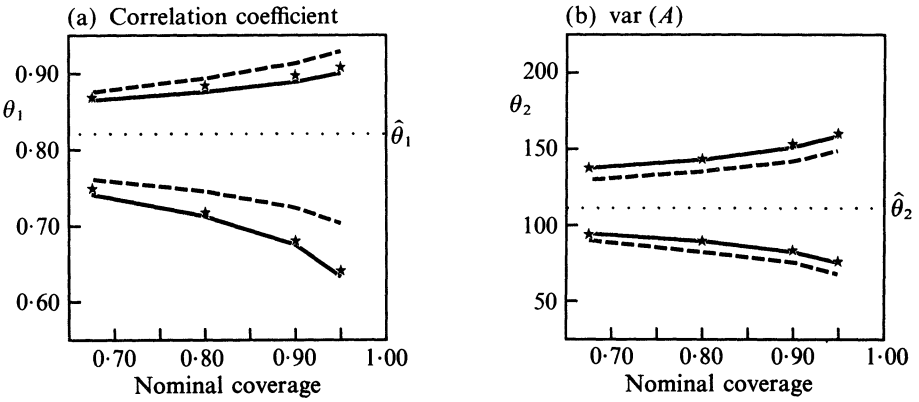

Fig. 3. Approximate nonparametric confidence intervals for the two parameters of Fig. 1, spatial test data; dashed lines are nonparametric standard intervals; solid lines are ABC intervals based on the empirical exponential family. Stars indicate nonparametric BC_a intervals, based on $N_B = 4800$ bootstrap replications.

the ABC algorithm, which requires no Monte Carlo simulations, gives an excellent approximation to the BC_a limits.

5. THEORETICAL JUSTIFICATION

Suppose that the assumed family of densities for an observed random variable $y = (y_1, \ldots, y_p)'$ is indexed by an unknown vector parameter $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_p)'$ and that the log likelihood function for $\bar{\eta}$ based on y has the exponential family form $l(\bar{\eta}; y) = n\{\bar{\eta}'y - \bar{\psi}(\bar{\eta})\}$, where $y - E(y)$ is $O_p(n^{-\frac{1}{2}})$ and both $\bar{\eta}$ and $\bar{\psi}(\bar{\eta})$ are $O(1)$. Upon defining $\eta = n\bar{\eta}$ and $\psi(\eta) = n\bar{\psi}(\bar{\eta})$, the log likelihood function for η is $l(\eta; y) = \eta'y - \psi(\eta)$, which agrees with (2·1). The cumulant generating function for y is

$$\log E(e^{\xi'y}) = \psi(\eta + \xi) - \psi(\eta), \quad \xi = (\xi_1, \ldots, \xi_p)',$$

and hence the first three joint cumulants of $y_1, \ldots, y_p$ are

$$E(y_i) = \psi_i(\eta), \quad \operatorname{cov}(y_i, y_j) = \psi_{ij}(\eta), \quad \operatorname{cum}(y_i, y_j, y_k) = \psi_{ijk}(\eta) \quad (i, j, k = 1, \ldots, p),$$

where $\psi_i(\eta) = \partial\psi(\eta)/\partial\eta_i$, $\psi_{ij}(\eta) = \partial^2\psi(\eta)/\partial\eta_i\partial\eta_j$, etc. The rth-order joint cumulants are readily seen to be $O(n^{1-r})$, since $\psi(\eta) = n\bar{\psi}(\eta/n)$. Instead of working with the natural parameter η, it is sometimes convenient to use the expectation parameter $\mu = (\mu_1, \ldots, \mu_p)'$ which has components $\mu_i = E(y_i) = \psi_i(\eta)$ $(i = 1, \ldots, p)$. The maximum likelihood equations $y_i = \psi_i(\hat{\eta})$ $(i = 1, \ldots, p)$ yield $\hat{\mu} = y$.

Now suppose that constructing confidence limits for a real-valued parameter $\theta = t(\mu)$ is of interest. Let

$$t_i(\mu) = \partial t(\mu)/\partial\mu_i, \quad t_{ij}(\mu) = \partial^2 t(\mu)/\partial\mu_i\partial\mu_j \quad (i, j = 1, \ldots, p),$$

and put $\hat{t}_i = t_i(\hat{\mu})$, $\hat{t}_{ij} = t_{ij}(\hat{\mu})$. To facilitate notation, we adopt the standard convention whereby summation is assumed over all indices that appear twice in any expression, the range of summation being $1, \ldots, p$. Thus, the Cramér–Rao lower bound (2·4) is $\sigma = (\psi_{ij}t_it_j)^{\frac{1}{2}}$, and its estimator (2·5) is $\hat{\sigma} = (\hat{\psi}_{ij}\hat{t}_i\hat{t}_j)^{\frac{1}{2}}$, with $\hat{\psi}_{ij} = \psi_{ij}(\hat{\eta})$ $(i, j = 1, \ldots, p)$. Note that σ is $O(n^{-\frac{1}{2}})$ and $\hat{\sigma}$ is $O_p(n^{-\frac{1}{2}})$. The variance of the maximum likelihood estimator $\hat{\theta} = t(\hat{\mu})$ is $\sigma^2 + O(n^{-2})$.

If the distribution of $\hat{\theta}$ is continuous, then exact confidence limits for θ can be defined, at least theoretically, in terms of the distribution of the approximately pivotal quantity $V = (\hat{\theta} - \theta)/\hat{\sigma}$. Define $K_\eta(x) = \text{pr}\,(V \leqslant x;\ \eta)$. The point

$$\hat{\theta}_{\text{ex}}[1-\alpha] = \hat{\theta} - \hat{\sigma}K_\eta^{-1}(\alpha) \tag{5·1}$$

is an exact upper $1-\alpha$ confidence limit for θ in the sense that $\text{pr}\,(\theta \leqslant \hat{\theta}_{\text{ex}}[1-\alpha];\ \eta) = 1-\alpha$. In most situations (5·1) cannot be calculated explicitly because it depends on the unknown parameter η through $K_\eta^{-1}(\alpha)$. Nonetheless, a useful notion of second-order correctness for approximate confidence limits can be defined in terms of $\hat{\theta}_{\text{ex}}[1-\alpha]$.

An approximate upper $1-\alpha$ confidence limit $\hat{\theta}[1-\alpha]$ is said to be second-order correct if it differs from $\hat{\theta}_{\text{ex}}[1-\alpha]$ by $O_p(n^{-3/2})$. This definition follows the one given by Hall (1988). On the other hand, an approximate limit $\hat{\theta}[1-\alpha]$ is said to be second-order accurate if it has coverage error of order $O(n^{-1})$, that is if

$$\text{pr}\,(\theta \leqslant \hat{\theta}[1-\alpha];\ \eta) = 1 - \alpha + O(n^{-1}).$$

An easy argument shows that second-order correctness generally implies second-order accuracy.

Second-order correctness can often be established by using an expansion of $\hat{\theta}_{\mathrm{ex}}[1-\alpha]$ which we now develop. Standard calculations show that, with error of order $O(n^{-1})$,

$$E(V) = -\tfrac{1}{2}m_1 + \tfrac{1}{2}m_2 - m_3, \quad \mathrm{var}\,(V) = 1, \quad \mathrm{skew}\,(V) = -2m_1 - 3m_3, \tag{5·2}$$

where

$$m_1 = \psi_{ijk}t_it_jt_k/\sigma^3, \quad m_2 = \psi_{ij}t_{ij}/\sigma, \quad m_3 = \psi_{ik}\psi_{jl}t_{ij}t_kt_l/\sigma^3.$$

Note that m_1, m_2 and m_3 are $O(n^{-\frac{1}{2}})$. Fourth- and higher-order cumulants of V are $O(n^{-1})$ or smaller. Use of (5·2) in a Cornish–Fisher expansion of $K_\eta^{-1}(\alpha)$ yields

$$-K_\eta^{-1}(\alpha) = z_\alpha + (\tfrac{1}{6}m_1 - \tfrac{1}{2}m_2 + \tfrac{1}{2}m_3) + (\tfrac{1}{3}m_1 + \tfrac{1}{2}m_3)z_\alpha^2 + O(n^{-1}), \tag{5·3}$$

z_α being the $1-\alpha$ quantile of the standard normal distribution. Now, put $\hat{m}_i = m_i(\hat{\eta})$ $(i = 1, 2, 3)$ and set

$$a = \tfrac{1}{6}\hat{m}_1, \quad z_0 = \tfrac{1}{6}\hat{m}_1 - \tfrac{1}{2}\hat{m}_2 + \tfrac{1}{2}\hat{m}_3,$$

so that $\hat{m}_i = m_i + O_p(n^{-1})$ $(i = 1, 2, 3)$ and both a and z_0 are $O_p(n^{-\frac{1}{2}})$. It follows from (5·3) that

$$\hat{\theta}_{\mathrm{ex}}[1-\alpha] = \hat{\theta} + \hat{\sigma}\{z_\alpha + z_0 + (2a + \tfrac{1}{2}\hat{m}_3)z_\alpha^2\} + O_p(n^{-3/2}). \tag{5·4}$$

If we also define $\hat{\gamma} = \frac{1}{2}(\hat{m}_2 - \hat{m}_3)$, then $z_0 = a - \hat{\gamma}$. When $p = 1$, m_2 and m_3 coincide, so $\hat{\gamma} = 0$ and $z_0 = a$. Similarly, $z_0 = a$ when θ is a linear combination of $\mu_1, \ldots, \mu_p$, since m_2 and m_3 both vanish in that case.

A drawback of definition (5·1) for the exact limit is its failure to transform appropriately under reparameterization. If another scalar parameter ϕ is considered such that $\phi = g(\theta)$ and g is monotonically increasing, then $\hat{\phi}_{\mathrm{ex}}[1-\alpha]$ and $g(\hat{\theta}_{\mathrm{ex}}[1-\alpha])$ are generally not the same. It can be shown from (5·4), however, that the difference $\hat{\phi}_{\mathrm{ex}}[1-\alpha] - g(\hat{\theta}_{\mathrm{ex}}[1-\alpha])$ is $O_p(n^{-3/2})$ in general, and hence the property of second-order correctness is preserved under parameter transformations.

Several methods are available for constructing second-order correct confidence limits. A simple approach is to use the approximation on the right-hand side of (5·4), which requires knowledge of $\hat{m}_1$, $\hat{m}_2$ and $\hat{m}_3$. Direct calculation of these estimators can be cumbersome, since it requires computing sums, up to fourth order, of partial derivatives, up to third order. For the ABC procedure, the quantities a, $\hat{b} = \frac{1}{2}\hat{\sigma}\hat{m}_2$ and $c_q = \frac{1}{2}\hat{m}_3$ are expressed in terms of simple derivatives that are evaluated numerically.

The derivatives involved in the ABC method are as follows. By definition,

$$\hat{t}_i = [dt(\hat{\mu} + he^i)/dh]_{h=0} \quad (i = 1, \ldots, p), \tag{5·5}$$

where e^i is the p-dimensional unit vector whose ith component is 1. The relationship $\mu_i = \psi_i(\eta)$ implies

$$\hat{\psi}_{ij} = [d\mu_i(\hat{\eta} + he^j)/dh]_{h=0} \quad (i, j = 1, \ldots, p), \tag{5·6}$$

and by setting $\hat{\sigma} = (\hat{\psi}_{ij}\hat{t}_i\hat{t}_j)^{\frac{1}{2}}$ and $\hat{\varepsilon} = (\hat{t}_1, \ldots, \hat{t}_p)'$,

$$a = \frac{1}{6}\frac{\hat{\psi}_{ijk}\hat{t}_i\hat{t}_j\hat{t}_k}{\hat{\sigma}^3} = \frac{1}{6}\frac{[d^2\hat{t}_i\mu_i(\hat{\eta} + h\hat{\varepsilon})/dh^2]_{h=0}}{\hat{\sigma}^3}. \tag{5·7}$$

Put $\hat{\boldsymbol{\psi}} = (\hat{\psi}_{ij})$, so that $\hat{\boldsymbol{\psi}}$ is a $p \times p$ matrix. Take D to be a diagonal matrix of eigenvalues of $\hat{\boldsymbol{\psi}}$ and Γ to be an orthogonal matrix whose columns consist of corresponding eigenvectors. In particular, let d_i be the ith diagonal element of D, and let $\gamma^i = (\gamma^i_1, \ldots, \gamma^i_p)'$ be

the ith column of Γ. Since $\hat{\ddot{\ell}} = \Gamma D \Gamma'$, it follows that

$$\hat{\psi}_{ij} = \sum_{k=1}^{p} d_k \gamma_i^k \gamma_j^k \quad (i, j = 1, \ldots, p).$$

Then

$$\hat{b} = \tfrac{1}{2}\hat{\psi}_{ij}\hat{t}_{ij} = \tfrac{1}{2}\sum_{k=1}^{p} [d^2 t(\hat{\mu} + h d_k^{1/2} \gamma^k)/dh^2]_{h=0}. \tag{5·8}$$

Finally, with $\hat{\delta}_i = \hat{\psi}_{ij}\hat{t}_j$ $(i = 1, \ldots, p)$ and $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_p)'$,

$$c_q = \frac{1}{2}\frac{\hat{\psi}_{ik}\hat{\psi}_{jl}\hat{t}_{ij}\hat{t}_k\hat{t}_l}{\hat{\sigma}^3} = \frac{1}{2}\frac{[d^2 t(\hat{\mu} + h\hat{\delta}/\hat{\sigma})/dh^2]_{h=0}}{\hat{\sigma}}. \tag{5·9}$$

Having calculated a, $\hat{b}$ and c_q in this way, we then obtain $\hat{\gamma} = \hat{b}/\hat{\sigma} - c_q$ and $z_0 = a - \hat{\gamma}$. The ABC algorithm actually computes z_0 as $\Phi^{-1}\{2\Phi(a)\Phi(-\hat{\gamma})\}$; this quantity differs from $a - \hat{\gamma}$ by $O_p(n^{-1})$, and hence it can replace $a - \hat{\gamma}$ in (5·4) without affecting the $O_p(n^{-3/2})$ order of error.

The second-order correctness of the ABC confidence limits (2·6) and (2·7) is easily verified. These limits involve the quantity λ, which satisfies $\lambda = z_\alpha + z_0 + 2az_\alpha^2 + O_p(n^{-1})$. Therefore, the quadratic ABC limit (2·6) has the expansion

$$\hat{\theta}_{\mathrm{ABC}_q}[1-\alpha] = \hat{\theta} + \hat{\sigma}\{z_\alpha + z_0 + (2a + c_q)z_\alpha^2\} + O_p(n^{-3/2}). \tag{5·10}$$

Similarly, expansion of the ABC limit (2·7) yields

$$\begin{aligned}\hat{\theta}_{\mathrm{ABC}}[1-\alpha] &= t(\hat{\mu}) + \frac{\hat{t}_i\hat{\delta}_i}{\hat{\sigma}}\lambda + \frac{1}{2}\frac{\hat{t}_{ij}\hat{\delta}_i\hat{\delta}_j}{\hat{\sigma}^2}\lambda^2 + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}\{z_\alpha + z_0 + (2a + c_q)z_\alpha^2\} + O_p(n^{-3/2}),\end{aligned} \tag{5·11}$$

as $\hat{t}_i\hat{\delta}_i = \hat{\sigma}^2$ and $\frac{1}{2}\hat{\sigma}^{-2}\hat{t}_{ij}\hat{\delta}_i\hat{\delta}_j = \hat{\sigma}c_q$. Since $c_q = \frac{1}{2}\hat{m}_3$, comparison of (5·10) and (5·11) with (5·4) shows that $\hat{\theta}_{\mathrm{ABC}_q}[1-\alpha]$ and $\hat{\theta}_{\mathrm{ABC}}[1-\alpha]$ both differ from $\hat{\theta}_{\mathrm{ex}}[1-\alpha]$ by $O_p(n^{-3/2})$.

In the present context, the parametric BC_a upper $1-\alpha$ confidence limit is

$$\hat{\theta}_{\mathrm{BC}_a}[1-\alpha] = \hat{G}^{-1}[\Phi\{z_0 + w(1-aw)^{-1}\}],$$

where $w = z_0 + z_\alpha$ and $\hat{G}$ is the parametric bootstrap distribution function, that is, $\hat{G}(s) = \mathrm{pr}\{t(y^*) \leqslant s\}$ for y^* distributed according to density (2·1) with $\eta = \hat{\eta}$. To demonstrate the second-order correctness of $\hat{\theta}_{\mathrm{BC}_a}[1-\alpha]$, let H_η be the distribution function of $U = (\hat{\theta} - \theta)/\sigma$. With error of order $O(n^{-1})$,

$$E(U) = \tfrac{1}{2}m_2, \quad \mathrm{var}(U) = 1, \quad \mathrm{skew}(U) = m_1 + 3m_3;$$

fourth- and higher-order cumulants of U are $O(n^{-1})$ or smaller. Thus we have the Cornish–Fisher expansion

$$-H_\eta^{-1}(\alpha) = z_\alpha + (\tfrac{1}{6}m_1 - \tfrac{1}{2}m_2 + \tfrac{1}{2}m_3) - (\tfrac{1}{6}m_1 + \tfrac{1}{2}m_3)z_\alpha^2 + O(n^{-1}).$$

Moreover,

$$\hat{G}^{-1}(\alpha) = \hat{\theta} - \hat{\sigma}\{z_\alpha + z_0 - (a + \tfrac{1}{2}\hat{m}_3)z_\alpha^2\} + O_p(n^{-3/2}),$$

since $\hat{G}^{-1}(\alpha) = \hat{\theta} + \hat{\sigma}H_\eta^{-1}(\alpha)$ by definition. Note that

$$z_0 + w(1-aw)^{-1} = z_\alpha + 2z_0 + az_\alpha^2 + O_p(n^{-1}),$$

and recall $-z_\alpha = \Phi^{-1}(1-\alpha)$. Hence,

$$\begin{aligned}\hat{\theta}_{\mathrm{BC}_a}[1-\alpha] &= \hat{G}^{-1}\{\Phi(z_\alpha + 2z_0 + az_\alpha^2)\} + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}\{z_\alpha + 2z_0 + az_\alpha^2 - z_0 + (a + \tfrac{1}{2}\hat{m}_3)z_\alpha^2\} + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}\{z_\alpha + z_0 + (2a + \tfrac{1}{2}\hat{m}_3)z_\alpha^2\} + O_p(n^{-3/2}),\end{aligned}$$

so $\hat{\theta}_{\mathrm{BC}_a}[1-\alpha]$ differs from $\hat{\theta}_{\mathrm{ex}}[1-\alpha]$ by $O_p(n^{-3/2})$.

Second-order correctness of BC_a confidence limits is discussed generally for parametric and nonparametric situations by Hall (1988) and by T. DiCiccio and J. P. Romano in an unpublished technical report.

References

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

Barndorff-Nielsen, O. E. & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. R. Statist. Soc.* B **41**, 279–312.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language.* Pacific Grove, California: Wadsworth and Brooks/Cole.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B **49**, 1–39.

Cox, D. R. & Snell, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

DiCiccio, T. J. & Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *J. Am. Statist. Assoc.* **82**, 163–70.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6**, 362–76.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72**, 45–58.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Am. Statist. Assoc.* **82**, 171–200.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16**, 927–85.

Konishi, S. (1991). Normalizing transformations and bootstrap confidence intervals. *Ann. Statist.* **19**, 2209–25.

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

McCullagh, P. (1984). Local sufficiency. *Biometrika* **71**, 233–44.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proc. 3rd Berkeley Symp.* **1**, 187–95. Berkeley: University of California Press.

Welch, B. L. & Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B **25**, 318–29.

Withers, C. S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. *Ann. Statist.* **11**, 577–87.

[*Received March* 1991. *Revised October* 1991]