BETTER APPROXIMATE CONFIDENCE INTERVALS
IN EXPONENTIAL FAMILIES

BY

T. DICICCIO and B. EFRON

TECHNICAL REPORT NO. 137

June 1990

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215–15

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Better Approximate Confidence Intervals in Exponential Families

By

T. DiCiccio and B. Efron

Technical Report No. 137

June 1990

Prepared Under the Auspices

Of

Public Health Service Grant 2 R01 GM21215-15

Also supported by National Science Foundation Grant DMS89-05874 and issued as Technical Report No. 345, Department of Statistics, Stanford University.

Division of Biostatistics

Stanford University

Stanford, California
Better Approximate Confidence Intervals in Exponential Families

T. DiCiccio and B. Efron

Abstract

Bootstrap confidence intervals usually required massive amounts of computation, on the order of 2,000 recomputations of the statistic of interest. However, most of this work can be avoided in sufficiently regular exponential families. In this article, a simple algorithm is developed to construct accurate confidence limits (called ABC limits) for scalar parameters in exponential families. This algorithm effectively approximates the \( BC_a \) bootstrap confidence limits and requires about 1% as much computation. It produces limits that have coverage error of order \( O(n^{-1}) \), where \( n \) is sample size, thereby improving on the limits obtained from standard normal theory which have coverage error of order \( O(n^{-1/2}) \). Higher-order correction terms, such as the bias correction and the acceleration constant required for the \( BC_a \) method, are computed in the algorithm by straightforward numerical differentiation. The ABC limits agree to second-order with ones obtained from profile likelihood, conditional likelihood, conditional profile likelihood, and Bayesian procedures. In particular, when the parameter of interest is a component of the natural parameter and conditional inference is appropriate, the ABC limits have coverage error of order \( O(n^{-1}) \) conditionally. Use of the algorithm is illustrated in examples concerning logistic regression, contingency tables, and multivariate analysis, among others. The Appendix contains a computer program for the algorithm written in the language \( S \).

Key Words: \( BC_a \) method; Conditional inference; Confidence limit; Coverage accuracy; Likelihood ratio statistic; Natural parameter; Parametric bootstrap; Second-order correct.
Better Approximate Confidence Intervals in Exponential Families

T. DiCiccio and B. Efron

1. Introduction.

Most parametric data analyses involve familiar exponential families: normal, binomial, Poisson, gamma, multinomial, Gaussian regression and ANOVA models, logistic regression, contingency tables, log linear models, multivariate normal problems, etc. Usually there will be a parameter of particular interest to the statistician, or perhaps several such parameters, for which a confidence interval is required. This paper presents a numerical algorithm for calculating highly accurate approximate intervals. The algorithm is based on bootstrap ideas but uses analytic approximations to avoid Monte Carlo sampling. As a result, it requires only a small fraction of the computer time needed for full bootstrap methods, on the order of a few percent.

Figure 1 concerns the lawschool data (shown), comprising pairs of achievement measures \((\text{LSAT}_i, \text{GPA}_i)\) for the 1975 entering classes at \(n = 15\) American law schools. Two parameters are considered, the Pearson correlation coefficient between LSAT and GPA, and the variance of the LSAT score. These have maximum likelihood estimates (MLEs) \(\hat{\theta} = .776\) and \(\hat{\sigma} = 1630\) respectively, assuming as we shall here that the pairs \((\text{LSAT}, \text{GPA})\) are a random sample from a bivariate normal distribution.

The dashed curves indicate the standard intervals

\[
\theta \in \hat{\theta} \pm z_\alpha \hat{\sigma},
\]

where \(z_\alpha\) is the \(100 \cdot (1 - \alpha)\)th percentile point of a standard normal distribution, \(z_{.05} = 1.645\) for example, and \(\hat{\sigma}\) is the usual estimate of standard error for \(\hat{\theta}\) obtained by differentiation of the log likelihood function. Here \(\hat{\sigma} = .103\) for the correlation coefficient, and \(\hat{\sigma} = 595\) for \(\text{var}\{\text{LSAT}\}\). Interval (1.1) has nominal coverage probability \(1 - 2\alpha\). This is usually a good estimate of the overall two-sided coverage probability, but can be quite misleading for the one-sided errors. In the correlation case for example, the standard intervals are much too conservative at the upper limits, and much too liberal at the lower limits.

The solid curves in Figure 1 indicate the ABC intervals, which are the subject of this paper. (ABC abbreviates “approximate bootstrap confidence” interval, or “approximate BC\(\alpha\)” interval.) These each required about 3 times the amount of computation as the standard intervals. Unlike the standard intervals, which are symmetrically placed about the MLE \(\hat{\theta}\), the ABC intervals make substantially different inferences for the upper and lower endpoints. This results in a large shift downwards for the correlation intervals, and a large shift upwards in the case of \(\text{var}\{\text{LSAT}\}\).

Exact confidence intervals are shown for both parameters in Figure 1, based on the bivariate normal model. The exact limits, indicated by stars, show that the ABC intervals are close to being exactly correct here.
Figure 1. Exact and approximate confidence intervals for two parameters, law school data (shown at right), assuming a bivariate normal model. 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 limits, for two-sided intervals of coverage 68%, 80%, 90%, and 95%. Dotted line indicates maximum likelihood estimate $\hat{\theta}$.

The standard intervals are automatic, in the sense that the statistician does not need to make special theoretical calculations for each new application. In principle, a single computer algorithm gives (1.1) in any situation. All the statistician supplies is the data, the functional form of the likelihood function, and the parameter of interest. Almost the same is true of the ABC intervals. One additional input is required, the function relating the natural parameter to the expectation parameter of the relevant exponential family. This function, which is usually an easy one to write down, is given for several common exponential families in Section 3.

We begin in Section 2 with a description and discussion of the ABC algorithm. Section 3 shows how the algorithm applies to some familiar exponential families, especially the more elaborate ones like logistic regression models, the multivariate normal, and contingency tables. Section 4 presents theoretical justification for the ABC intervals, showing that they are second-order accurate. That is, they have asymptotic error probabilities $\alpha + O(1/n)$ in each tail, compared to $\alpha + O(1/\sqrt{n})$ for the standard intervals.

It is also shown that the ABC intervals are second-order correct: the interval endpoints differ from the exact endpoints by only $O_p(\hat{\sigma}/n)$, compared to $O_p(\hat{\sigma}/\sqrt{n})$ for the standard intervals. This assumes that exact intervals exist, as in Figure 1. In most problems exact intervals do not exist, but there are other arguments for the second-order correctness of the ABC intervals. In particular
in McCullagh (1984), Section 4 shows that the ABC intervals agree to second-order with those suggested by the theories of Barndorff-Nielsen (1986), Barndorff-Nielsen and Cox (1979), Cox and Reid (1987), Welch and Peers (1963), and Hall (1988), as well as the bootstrap $BC_a$ method, Efron (1987). These papers, and others by Konishi (1989), Withers (1983), and DiCiccio and Tibshirani (1987) involve theoretical calculations very much like the ones presented here. The main novelty of this paper is the emphasis on efficient, routine calculation of second-order intervals.

The appendix displays a program for implementing the ABC calculations, written in the language S, and discusses some computational details.

The presentation here proceeds at two levels: Section 3 and the appendix contain everything needed for the actual calculation of the ABC intervals. Sections 2 and 4 discuss the underlying theory, first heuristically and then giving more careful proofs.

2. ABC Intervals.

The ABC intervals are approximations to the bootstrap $BC_a$ intervals, Efron (1987), using analytic methods to avoid the $BC_a$’s Monte Carlo calculations. This Section describes the ABC algorithm, emphasizing the simple nature of the second order corrections it makes to the standard intervals (1.1). A detailed discussion of the second order asymptotics is given in Section 4. We begin with a brief description of the $BC_a$ method as it applies to parametric problems. Readers mainly interested in computation of the ABC intervals can go directly to Section 3 and the appendix, which are self-contained.

Suppose that the data vector $y$ has a family of possible density functions indexed by a parameter vector $\mu$. Let $\theta = t(\mu)$ indicate a real-valued parameter of particular interest, estimated by its MLE $\hat{\theta} = t(\hat{\mu})$ where $\hat{\mu}$ is the MLE of the entire vector $\mu$. The parametric bootstrap cumulative distribution function (CDF) is defined to be

$$\hat{G}(s) \equiv \text{Prob}_{\hat{\mu}}\{t(\hat{\mu}^*) < s\}, \quad (2.1)$$

where $\hat{\mu}^*$ is the MLE for $\mu$ based on a hypothetical observation $y^*$ distributed according to the density function indexed by $\mu = \hat{\mu}$.

The $BC_a$ intervals are calculated in terms of $\hat{G}$ and two derived constants, the acceleration $a$ and the bias-correction $z$, see Sections 2-4 of Efron (1987). Let

$$z[1 - \alpha] \equiv z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}. \quad (2.2)$$

The $BC_a$ interval of nominal coverage $1 - 2\alpha$ is defined to be

$$\theta \in [\hat{G}^{-1}\Phi(z[\alpha]), \hat{G}^{-1}\Phi(z[1 - \alpha])], \quad (2.3)$$

$\Phi$ being the standard normal cdf. This is second order accurate, Hall (1988), and second order correct, see Section 4. The only difficulty with (2.3) is computational: it requires perhaps 1000 to 4000 times as much computation as the original calculation of $\hat{\theta} = t(\hat{\mu})$. The ABC algorithm calculates simple analytic approximations for $z_0$, $\alpha$, and $\hat{G}^{-1}$, giving a good approximation to interval (2.3) at a small fraction of the computational expense.
We will assume that our parametric family of densities is an exponential family, with 
$p$-dimensional natural parameter vector $\eta$, so the density can be expressed as
\[ g_\eta(y) = e^{\eta' y - \psi(\eta)}. \] (2.4)
The $p$-dimensional data vector $y$ is usually a sufficient statistic obtained from a larger data set
involving $n$ components, as shown by the example that follows. The expectation parameter vector
$\mu \equiv E_\eta\{y\}$ can be obtained by differentiating $\psi(\eta),$
\[ \mu = \dot{\psi}(\eta) \equiv (\cdots\partial\psi(\eta)/\partial\eta_i\cdots)' . \] (2.5)
The expectation vector $\mu$ indexes the exponential family just as well as does $\eta$, since the mapping
from $\eta$ to $\mu$ 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 $\eta$ corresponding to $\mu$. Section 2 of Efron (1978) gives a brief
review of exponential families.

The $p \times p$ covariance matrix of $y$, denoted either $\Sigma(\eta)$ or $\Sigma(\mu)$, is given by the second derivative
matrix of $\psi,$
\[ \Sigma(\mu) = \ddot{\psi}(\eta) \equiv \left( \frac{\partial^2 \psi}{\partial\eta_i \partial\eta_j} \right) \quad i, j = 1, 2, \cdots, p. \] (2.6)
Likewise, third derivatives of $\psi(\eta)$ give third order cumulants,
\[ \frac{\partial^3 \psi(\eta)}{\partial\eta_i \partial\eta_j \partial\eta_k} = E_\eta\{(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k)\}, \] (2.7)
etc.

The MLE of the expectation vector $\mu$ is $y$, say $\hat{\mu} = y$, so the MLE 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
\[ \sigma = \left[ i(\hat{\mu})' \dot{\Sigma}(\mu) i(\hat{\mu}) \right]^{1/2}, \] (2.8)
this also being the Cramer-Rao lower bound (CRLB) for the standard deviation of an unbiased
estimator of $\theta$. In practice, $\hat{\mu} = y$ substitutes for $\mu$ in (2.8), giving an estimated standard error for
$\hat{\theta}$
\[ \hat{\sigma} = \left[ i(\hat{\mu})' \dot{\Sigma}(\hat{\mu}) \right]^{1/2}, \] (2.9)
where $\dot{\Sigma} \equiv \dot{\Sigma}(\hat{\mu}) = \dot{\Sigma}(y)$. It turns out that $\dot{\Sigma}$ is the inverse of the estimated Fisher information
matrix for $\hat{\mu}$, so expression (2.9) is the usual estimate of standard error for $\hat{\theta}$ based on differentiating
the log likelihood function.

As an example of our exponential family definitions, consider observing a random sample of
size $n$ from a distribution with density $(1/\mu)e^{-x/\mu}$ on $x > 0$, $\mu$ unknown,
\[ x_1, x_2, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mu \text{ Gamma}_1, \] (2.10)
Gamma$_1$ denoting a gamma distribution with shape parameter 1. This is a one-parameter ($p = 1$) exponential family, with the sufficient statistic $y = \bar{x} = \sum_{i=1}^{n} x_i/n$ having a scaled gamma
distribution with shape parameter $n$,
\[ y \sim \mu \text{ Gamma}_n/n \quad \left[ \text{density} \frac{n(ny)^{n-1}e^{-ny/\mu}}{\mu^n \Gamma(n)} \text{ for } y > 0 \right] \] (2.11)
The MLE \( \hat{\mu} \) equals \( y \).

Notice that \( \mu \) is the expectation parameter of \( y \) as before, \( \mu = E_\mu(y) \). Expressing density (2.10) in form (2.4) gives

\[
\eta = -n/\mu \quad \text{and} \quad \psi(\eta) = -n \log(\eta).
\] (2.12)

(Factors like \( n^ny^{-1}/\Gamma(n) \) that do not involve unknown parameters can be ignored in these calculations.) Differentiating \( \psi(\eta) \) as in (2.6), (2.7) gives \( \text{var}_\eta(y) = \mu^2/n, E_\eta((y - \mu)^2) = 2\mu^3/n^2 \), so \( y \) has standard deviation and skewness

\[
\sigma(\mu) \equiv \text{sd}_\mu(y) = \mu/\sqrt{n}, \quad \text{and} \quad \gamma(\mu) \equiv \text{skew}_\mu(y) = 2/\sqrt{n}.
\] (2.13)

The standard deviation \( \sigma(\mu) \) equals the Cramer-Rao lower bound for the unbiased estimation of \( \mu \), (2.8).

The ABC intervals can be motivated in terms of this example. Suppose we want a confidence interval for \( \mu \) in situation (2.10) or (2.11). In this case exact limits for \( \mu \) can be obtained from the usual confidence interval construction for a one-parameter family. Letting \( \hat{\mu}_{EX}[1 - \alpha] \) be the endpoint of the exact 100(1 - \( \alpha \))\textsuperscript{th} one-sided upper interval, we calculate

\[
\hat{\mu}_{EX}[1 - \alpha] = \frac{n}{\text{Gamma}_n(\alpha)} y,
\]

\( \text{Gamma}_n(\alpha) \) indicating the \( 100\alpha \)\textsuperscript{th} percentile of a standard gamma distribution with shape parameter \( n \). For example with \( n = 10 \) and \( \alpha = .05, \hat{\mu}_{EX}[.95] = (10/5.425)y = 1.843y \); the exact two-sided 90% interval for \( \mu \) is \([\hat{\mu}_{EX}[.05], \hat{\mu}_{EX}[.95]] = [.6367y, 1.843y] \).

The \( BC_\alpha \) interval for \( \mu \) is calculated from (2.1), (2.3). Here the parametric bootstrap takes \( y^* \sim y \text{ Gamma}_n/n \) and \( \hat{\mu}^* = y^* \), so \( \hat{G} \) is the cdf of the distribution \( y \text{ Gamma}_n/n \), giving endpoints

\[
\hat{\mu}_{BC_\alpha}[1 - \alpha] = \frac{\text{Gamma}(1-\tilde{\alpha})}{n} y,
\] (2.14)

with

\[
1 - \tilde{\alpha} = \Phi(z_0 + z_{0\alpha}), \quad \text{where} \quad z_{0\alpha} = \frac{z_0 + z_{\alpha}}{1 - \alpha(z_0 + z_{\alpha})},
\] (2.14a)

For a one-parameter exponential family, both \( z_0 \) and \( a \) approximately equal \( \text{skew}_\mu(y)/6 = 1/(3\sqrt{n}) \), Efron (1987). Table 1 shows close agreement between the exact and the \( BC_\alpha \) confidence limits when \( n = 10 \).

In this example there is an easy expression for the bootstrap cdf (2.1), which leads to the simple formula (2.14) for the \( BC_\alpha \) confidence limits. This does not usually happen, especially in nonparametric problems, so Monte Carlo simulations must be employed to calculate the \( BC_\alpha \) intervals.

The ABC algorithm avoids Monte Carlo by approximating the percentiles of the bootstrap cdf \( \hat{G} \) with a two-term Cornish-Fisher expression,

\[
\hat{G}^{-1}(1 - \alpha) \approx \hat{\mu} + \hat{\sigma}[z_{\alpha} + \frac{\hat{\sigma}^2}{6}(z_{\alpha}^2 - 1)],
\] (2.15)
where \( \mu, \sigma, \) and \( \gamma \) are the mean, standard deviation, and skewness of \( \hat{G} \). In our case, \( \hat{G} \) is the cdf of the distribution \( y \Gamma_{a/n} \), so \( \mu = y, \sigma = y/\sqrt{n}, \gamma = 2/\sqrt{n} \) (and \( \gamma/6 \) is the value of \( a \) used above). The ABC confidence limits are obtained using (2.15) to approximate (2.14) or (2.3),

\[
\hat{\mu}_{ABC}[1 - \alpha] = [1 + \frac{1}{\sqrt{n}} z_{1\alpha}] y
\]

(\text{where } z_{1\alpha} = z_{0\alpha} + a z_{0\alpha}^2). \tag{2.16}

Some higher-order terms have been dropped in going from (2.14) to (2.16).

Table 1 shows excellent agreement between \( \hat{\mu}_{ABC}, \hat{\mu}_{BC_a}, \) and \( \hat{\mu}_{EX} \). The endpoints of the standard interval,

\[
\hat{\mu}_{STAN}[1 - \alpha] = [1 + \frac{1}{\sqrt{n}} z_{\alpha}] y
\]

are considerably less accurate in this case.

We can extend example (2.11) to other exponential families. Suppose that \( g_\mu(y) = e^{y - \psi(y)} \) is any one-parameter exponential family of densities, with sufficient statistic \( y \) and expectation parameter \( \mu \). The ABC confidence limits for \( \mu \) are defined to be

\[
\hat{\mu}_{ABC}[1 - \alpha] = \hat{\mu} + \hat{\sigma}(\mu) z_{1\alpha}
\]  \tag{2.18}

where \( z_{1\alpha} \) is as given in (2.14a), (2.16), and

\[
\hat{\mu} = y, \quad \hat{\sigma}(\mu) = \bar{\psi}(\bar{\gamma})^{\frac{1}{2}}, \quad z_0 = a = \gamma/6 = \bar{\psi}(\bar{\gamma})/6\bar{\psi}(\bar{\gamma})^{3/2}. \tag{2.18a}
\]

As before, (2.18) is obtained by approximating the \( BC_a \) endpoints (2.3) with a two-term Cornish-Fisher expansion.

<table>
<thead>
<tr>
<th>Method</th>
<th>.05</th>
<th>.10</th>
<th>.16</th>
<th>.50</th>
<th>.84</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>10/\Gamma_{10}^{(a)}</td>
<td>0.6367</td>
<td>0.7039</td>
<td>0.7640</td>
<td>1.034</td>
<td>1.448</td>
<td>1.607</td>
</tr>
<tr>
<td>BC_a</td>
<td>Gamma_{10}^{(1-a)}/10^*</td>
<td>0.6365</td>
<td>0.7037</td>
<td>0.7638</td>
<td>1.034</td>
<td>1.448</td>
<td>1.607</td>
</tr>
<tr>
<td>ABC</td>
<td>( 1 + \frac{1}{\sqrt{10}} z_{1\alpha} )</td>
<td>0.6396</td>
<td>0.7056</td>
<td>0.7650</td>
<td>1.034</td>
<td>1.445</td>
<td>1.602</td>
</tr>
<tr>
<td>Standard</td>
<td>( 1 + \frac{1}{\sqrt{10}} z_{\alpha} )</td>
<td>0.4797</td>
<td>0.5947</td>
<td>0.6855</td>
<td>1.000</td>
<td>1.314</td>
<td>1.405</td>
</tr>
</tbody>
</table>

*see definitions (2.14a),(2.16)

**Table 1.** Exact and approximate confidence limits for \( \mu \), having observed \( y \sim \Gamma_{10}/10 \), (2.11); number tabulated is the indicated 100 \cdot (1 - \alpha) \text{th} \) confidence limit, divided by \( y \); constants \( a \) and \( z_0 \) equal skewness (\( \Gamma_{10} \))/6 = 1/(3\sqrt{10}) = .105. The second order corrections made to the standard intervals by the \( BC_a \) and ABC methods are quite large in this case, and give nearly the exact confidence limits.
Suppose that the parameter of interest is some monotone function of $\mu$, say $\theta = t(\mu)$, rather than $\mu$ itself. The ABC confidence limits for $\theta$ are defined to be the same function of (2.18),

$$\hat{\theta}_{ABC}[1 - \alpha] = t(\hat{\mu} + \hat{\sigma}(\mu)z_{1\alpha}) = t(\hat{\mu} + \frac{\hat{\sigma}(\theta)}{|d\theta/d\mu|} z_{1\alpha}).$$

(2.19)

Here $\hat{\sigma}(\theta)$ is the estimated standard error of $\hat{\theta}$, (2.9),

$$\hat{\sigma}(\theta) = \hat{\sigma}(\mu)|t(\hat{\mu})| = \hat{\nu}(\hat{\eta})^{\frac{1}{2}}|dt/d\mu|.$$

(2.20)

We now return to the multiparameter situation, where $y$ is observed from a $p$-dimensional exponential family (2.4), and we want to construct a confidence interval for a real-valued parameter of interest $\theta = t(\mu)$, $\mu$ being the expectation parameter vector (2.5). The ABC limit for $\theta$ is defined to be

$$\hat{\theta}_{ABC}[1 - \alpha] = t(\hat{\mu} + \frac{\hat{\delta}}{\hat{\sigma}} z_{1\alpha}),$$

(2.21)

where $\hat{\mu} = y$, $\hat{\sigma} = [t(\hat{\mu})^T \frac{\partial}{\partial \hat{\mu}} t(\hat{\mu})]^{\frac{1}{2}}$ as in (2.9), and $\hat{\delta}$ is the least favorable direction vector

$$\hat{\delta} \equiv \hat{\Sigma} t(\hat{\mu}).$$

(2.22)

The constant $z_{1\alpha}$ is defined as before,

$$z_{1\alpha} = z_{0\alpha} + az_{0\alpha}^2, \quad \text{where} \quad z_{0\alpha} = \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}.$$  

(2.23)

This bias constant $z_0$ may now differ from the acceleration constant $a$; the ABC algorithm calculates $z_0$ and $a$, as well as $\hat{\delta}/\hat{\sigma}$, by numerical differentiation, as described below. In the case $p = 1$, $\hat{\delta}/\hat{\sigma}$ equals $\hat{\sigma}(\mu)$, and (2.21) reduces to (2.19).

A linear Taylor series expansion of (2.21) gives

$$t(\hat{\mu} + \frac{\hat{\delta}}{\hat{\sigma}} z_{1\alpha}) \approx t(\hat{\mu}) + t(\hat{\mu})^T \frac{\partial}{\partial \hat{\mu}} \hat{\delta} z_{1\alpha},$$

(2.24)

$$= \hat{\theta} + \hat{\sigma} z_{1\alpha},$$

where we have used $t(\hat{\mu}) = \hat{\theta}$, and $t(\hat{\mu})^T \frac{\partial}{\partial \hat{\mu}} t(\hat{\mu}) = \hat{\sigma}^2$. If the constants $z_0$ and $a$ both equal zero, then $z_{1\alpha} = z_\alpha$, and the last expression becomes the standard interval limit

$$\hat{\theta}_{STAN}[1 - \alpha] = \hat{\theta} + \hat{\sigma} z_\alpha.$$  

(2.25)

In other words, $\hat{\theta}_{STAN}[1 - \alpha]$ equals $\hat{\theta}_{ABC}[1 - \alpha]$ if $z_0 = a = 0$ and $t(\mu)$ is linear.

We see that in general the ABC algorithm makes three second-order corrections to the standard intervals:

- **Non-linearity correction.** Adding the next term to Taylor series (2.24) gives

$$t(\hat{\mu} + \frac{\hat{\delta}}{\hat{\sigma}} z_{1\alpha}) \approx t(\hat{\mu}) + \hat{\sigma} z_{1\alpha} + \frac{1}{2} \frac{\hat{\delta}^T \hat{\delta}}{\hat{\sigma}^2} z_{1\alpha}^2,$$

(2.26)
where \( \tilde{t} \) is the \( p \times p \) matrix with elements \( \partial^2 t(\hat{\mu}) / \partial \hat{\mu}_i \partial \hat{\mu}_j \). The last term in (2.26), which is ignored in (2.25), is second order if \( t \) is nonlinear, that is it has magnitude \( O_p(\sqrt{n}) \) in repeated sampling situations like (2.10). (Sometimes it is advantageous to evaluate the ABC limits from the right side of (2.25) rather than the left, as discussed in Section 3.)

- **Bias correction.** In general, the MLE \( \hat{\theta} \) has bias \( O_p(\sqrt{n}) \) for estimating \( \theta \). This enters \( \hat{\theta}_{ABC}[1 - \alpha] \) via the bias-correction constant \( z_0 \) in definition (2.23), which makes \( z_{1\alpha} \) in (2.21) differ from \( z_\alpha \) in (2.25) by \( O_p(1/\sqrt{n}) \). In one-parameter problems \( z_0 \) equals the acceleration constant \( a \), (2.18a), but in multiparameter exponential families \( z_0 \) also depends on the curvature of the \( (p - 1) \)-dimensional level surface \( \{ t(\mu) = t(\hat{\mu}) \} \). We will see that about half of the numerical work in computing ABC intervals comes in the evaluation of this curvature.

- **Acceleration correction, and nonnormality.** Using the standard interval amounts to taking literally the first order asymptotic normal approximation,

\[
\frac{\hat{\theta} - \theta}{\hat{\sigma}} \sim N(0, 1).
\] (2.27)

Even in a simple situation like (2.11), the MLE can be nonnormal at the second order, with its cdf differing from \( \Phi \) by amount \( O(1/\sqrt{n}) \). As discussed in Section 2 of Efron (1987), normalizing \( \hat{\theta} \) by a transformation \( \hat{\phi} = h(\hat{\theta}) \), will in general lead to \( \hat{\phi} \) having non-constant standard deviation as a function of \( \phi \),

\[
\frac{d\sigma(\phi)}{d\phi} = a.
\] (2.28)

The acceleration constant \( a \) is of magnitude \( O(\sigma(\theta)/\sqrt{n}) \). In exponential families, the acceleration is calculated from the skewness of \( y \) along the direction \( t(\hat{\mu}) \), as discussed below, and can be calculated in terms of a single numerical second derivative. Then the value of \( a = \gamma / 6 \) is incorporated into \( z_{1\alpha} \) as in (2.23). This amounts to a correction both for the acceleration effect and for the nonnormality of \( \hat{\theta} \).

Figure 2 concerns the numerical calculation of the constants \( a \) and \( z_0 \). Stein (1956) originated the idea of reducing a multiparameter estimation problem to an equivalent one-parameter problem. In an exponential family, Stein's least favorable family is the one-parameter subfamily of (2.4) passing through \( \hat{\eta} \), the MLE of \( \eta \), in the direction \( \hat{\epsilon} \equiv t(\hat{\mu}) \), say

\[
\hat{g}_\lambda(y^*) = g_{\hat{\eta} + \lambda \hat{\epsilon}}(y^*) \quad [\hat{\epsilon} = t(\hat{\mu})].
\] (2.29)

This is illustrated on the left side of Figure 2. Here \( \hat{\eta} \) and \( \hat{\epsilon} \) are considered fixed, and only the scaler \( \lambda \) is variable. Using \( y^* \) for a hypothetical observation from \( \hat{g}_\lambda \) avoids confusion with the actual data vector \( y \), which gave \( \hat{\eta} \) and \( \hat{\epsilon} \).

It turns out to be just as difficult to estimate \( \theta = t(\mu) \) in the one parameter family \( \hat{g}_\lambda \) as in the full \( p \)-parameter family \( g_\eta \), which accounts for the name "least favorable". More precisely, the Cramer-Rao lower bound for estimating \( \theta \) in \( \hat{g}_\lambda \), evaluated at \( \lambda = 0 \), equals \( \hat{\sigma} = [t(\hat{\mu})^T \hat{\Sigma} t(\hat{\mu})]^{1/2} \), (2.9), the Cramer-Rao lower bound for estimating \( \theta \) in \( g_\eta \), evaluated at \( \eta = \hat{\eta} \).

The acceleration constant \( a \) used in the ABC construction is calculated from the skewness of \( \hat{g}_\lambda \) in the direction \( \hat{\epsilon} \) at \( \lambda = 0 \):
Figure 2. Stein's least favorable family passes through \( \hat{\eta} \) in the direction \( \hat{\epsilon} = \hat{t}(\hat{\mu}) \) in the natural parameter space; it passes through \( \hat{\mu} = y \) in the direction \( \hat{\delta} = \hat{v} \hat{t}(\hat{\mu}) \) in the expectation parameter space; the standard error \( \hat{\sigma} \) and acceleration constant \( a \) are determined by the least favorable family; the bias-correction constant \( x_0 \) depends also on the curvature of the level surface \( \{ t(\mu) = t(\hat{\mu}) \} \) at \( \hat{\mu} = y \).

\[
a = \text{skew}_{\lambda=0}(\hat{\epsilon}'y^*)/6 = \frac{\partial^3}{\partial \lambda^3} \psi(\hat{\eta} + \lambda \hat{\epsilon})|_{\lambda=0}/6 \cdot \left( \frac{\partial^2}{\partial \lambda^2} \psi(\hat{\eta} + \lambda \hat{\epsilon})|_{\lambda=0} \right)^{3/2}
\]

(2.30)

where \( \mu(\hat{\eta} + \lambda \hat{\epsilon}) \) is the expectation vector corresponding to natural parameter \( \hat{\eta} + \lambda \hat{\epsilon} \). The last line of (2.29) is based on two exponential family relationships:

\[
\frac{\partial \psi(\hat{\eta} + \lambda \hat{\epsilon})}{\partial \lambda} = \hat{\epsilon}' \mu(\hat{\eta} + \lambda \hat{\epsilon}), \quad \text{and} \quad \frac{\partial^2 \psi(\hat{\eta} + \lambda \hat{\epsilon})}{\partial \lambda^2}|_{\lambda=0} = \hat{\epsilon}' \hat{\psi}(\hat{\eta} + \lambda \hat{\epsilon}) \hat{\epsilon}|_{\lambda=0} = a^2.
\]

Section 6 of Efron (1987) discusses formula (2.30).

The least favorable family determines \( \hat{\sigma} \) (and hence the standard interval (1.1)) and also the acceleration constant \( a \). However there is one element of the ABC algorithm which cannot be obtained from the least favorable family: the curvature of the level surface \( \{ t(\mu) = t(\hat{\mu}) \} \) evaluated at \( \hat{\mu} = y \). It is this curvature which makes the bias-correction constant \( x_0 \) not equal the acceleration \( a \) in multi-parameter families. The right side of Figure 2 illustrates the situation in the expectation parameter space. The least favorable family (2.27) passes through \( \hat{\mu} = y \) in the least favorable direction \( \hat{\delta} = \hat{v} \hat{\epsilon} = \hat{v} \hat{t}(\hat{\mu}) \), (2.22). The \((p - 1)\)-dimensional level surface \( \{ t(\mu) = t(\hat{\mu}) \} \) also passes through \( t(\hat{\mu}) \), orthogonally to \( \hat{\delta} \) in the metric \( \hat{\Sigma}^{-1} \). The total curvature \( \text{curv}(\hat{\mu}) \) of \( \{ t(\mu) = t(\hat{\mu}) \} \) at \( \hat{\mu} \), in the metric \( \hat{\Sigma}^{-1} \), is computed as follows: let \( C \) be any \( p \times p \) square root of \( \hat{\Sigma} \), for instance.
\[ C = \Gamma D^{1/2} \] from the orthogonal decomposition \( \hat{\Sigma} = \Gamma D \Gamma' \), with columns \( C_1, C_2, \ldots, C_p \). Then
\[
\text{curv}(\hat{\mu}) = \left\{ \frac{\sum_{j=1}^{p} \frac{\partial^2}{\partial \lambda^2} t(\hat{\mu} + \lambda C_j)|_{\lambda=0} - \frac{\partial^2}{\partial \lambda^2} t(\hat{\mu} + \lambda \hat{\sigma}/\hat{\sigma})|_{\lambda=0}}{2\hat{\sigma}} \right\}.
\] (2.31)

(This expression is the quantity called \( t(\lambda) \) in Theorem 1 of Efron (1985).) Formula (2.31) computes the sum of \( p - 1 \) orthogonal curvatures by first computing the sum of \( p \) curvatures in a convenient coordinate system, and then subtracting the curvature along the direction \( \hat{\sigma} \).

Finally, the bias-correction constant \( z_0 \) in the ABC algorithm (2.21)-(2.23) is calculated according to formula (8.8) of Efron (1987),
\[
z_0 = \Phi^{-1}\{2\Phi(a)\Phi(-\text{curv}(\hat{\mu}))\}, \tag{2.32}
\]
where \( \Phi \) is the standard normal cdf. The simpler formula
\[
z_0 = a - \text{curv}(\hat{\mu}) \tag{2.33}
\]
agrees with (2.32) to second-order. Notice that if \( \{t(\mu) = t(\hat{\mu})\} \) is flat, then \( \text{curv}(\hat{\mu}) = 0 \), so \( z_0 = a \) as in the one-dimensional case.

Despite the many approximations involved, the ABC intervals closely matched the full bootstrap \( BC_\alpha \) intervals in the cases we considered. As an example, Figure 3 compares the two methods in the following situation: we observe a random sample of size \( n \) from a scaled gamma distribution with scale and shape parameter unknown, \( x_1, x_2, \ldots, x_n \sim \alpha \text{ Gamma}_\nu \); we want confidence intervals for the parameter \( \mu_1 = \alpha \nu = E_{\alpha,\nu}\{x_i\} \) and also for \( \mu_2 = \log(\alpha) + dT(\nu)/d\nu = E_{\alpha,\nu}\{\log(x_i)\} \). This is a two-parameter exponential family, with sufficient statistic \( (y_1, y_2) = (\Sigma x_i, \Sigma \log(x_i))/n \). The case illustrated in Figure 3 is based on a sample with \( n = 10, \alpha = 1, \nu = 5, \) and \( y = (6.132, 1.732) \). Figure 3 shows excellent agreement between the ABC intervals and the \( BC_\alpha \) intervals based on \( B = 4000 \) Monte Carlo replications.

In summary, here are the computational steps required to calculate the ABC endpoints (2.21):

- The computation begins with the sufficient statistic \( y = \hat{\mu} \), the estimated covariance matrix \( \hat{\Sigma} = \Sigma(\hat{\mu}) \), and the MLE \( \hat{\eta} \) of the natural parameter \( \eta \). Section 3 gives \( \hat{\Sigma} \) and \( \hat{\eta} \) for the common exponential families, and also the function \( \mu(n) \) required in the ABC algorithm.
- Numerical differentiation gives \( \hat{\varepsilon} = \hat{\varepsilon}(\hat{\mu}) = (\cdots \partial t/\partial y_j \cdots)' \), then \( \hat{\delta} = \hat{\Sigma}(\hat{\mu})^{-1} \hat{\varepsilon}(\hat{\mu}) \), (2.22), and \( \hat{\sigma} = [\hat{\varepsilon}(\hat{\mu})' \hat{\Sigma}(\hat{\mu})^{-1} \hat{\varepsilon}(\hat{\mu})]^{1/2} \), (2.9).
- The acceleration constant \( a \) is calculated by numerical second differentiation, (2.30).
- The curvature \( \text{curv}(\hat{\mu}) \) is calculated from \( p + 1 \) numerical second derivatives, as in (2.31).
- The bias constant \( z_0 \) is calculated from (2.31), giving \( z_{10} \) (2.23).
- Finally, \( \hat{\theta}_{ABC}[1 - \alpha] \) is calculated as in (2.21).

We see that the ABC algorithm requires the calculation of \( p \) first derivatives and \( p + 2 \) second derivatives. The \( S \) program given in the Appendix computes each of these by evaluating \( t(y + \Delta) \) and \( t(y - \Delta) \) for a certain small value of the incremental vector \( \Delta \), for a total of \( 4p + 6 \) reevaluations of \( t \), including the final two at step (2.21). By comparison, the corresponding program to numerically
Figure 3. Approximate confidence intervals for $E(x)$, left panel, and $E(\log(x))$, right panel; sample of size 10 from $\alpha \Gamma \gamma \omega$ distribution, with $\alpha$ and $\nu$ unknown. The ABC interval, solid curve, agree well with the $BC_{\alpha}$ limits based on $B = 4000$ bootstrap replications, indicated by stars.

evaluate the standard intervals (1.1) requires $2p$ reevaluations of $t$ (in order to compute $i(\hat{\mu})$ for $\hat{\sigma}$, (2.9)), while the full bootstrap $BC_{\alpha}$ intervals require 1000 to 4000 reevaluations, though this total can often be reduced, see Efron (1990).

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 authors are investigating extensions of the ABC algorithm to non-exponential parametric families, and also to intervals of third order accuracy.

3. Familiar Exponential Families.

This section describes exactly how to call the $S$ program “abc” listed in the appendix, in order to calculate ABC confidence limits for parameters in the most widely-used exponential families: Poisson, multinomial, binomial, logistic regression, multivariate normal, and gamma.

The program call is of the form

$$abc(tt, y, S, etahat, mu, \cdots)$$

(3.1)
where \( tt \) is the function of the expectation vector \( \mu = E\{y\} \) for which a confidence interval is desired (previously referred to as \( t(\mu) \)); \( y \) is the sufficient vector of the exponential family; \( S \) is the maximum likelihood estimated covariance matrix of \( y \) (previously called \( \Sigma = \Sigma(\hat{\mu}) \)); \( \hat{\eta} \) equals \( \hat{\eta} \), the MLE of the natural parameter vector of the exponential family; and \( \mu \) is the function giving \( \mu \) in terms of \( \eta \) (the function \( \mu = \psi(\eta) \) in (2.5)). We will describe \( y, S, \hat{\eta}, \) and \( \mu \) for several familiar exponential families.

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

\[
y_j \sim Po(\mu_j) \quad \text{independently} \quad j = 1, 2, \ldots, p. \tag{3.2}
\]

In this case the data vector \( y = (y_1, y_2, \ldots, y_p)' \) has mean vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_p)' \), and covariance matrix \( \text{diag}(\mu) = \mu \times p \) diagonal matrix with jth diagonal element \( \mu_j \), so \( S = \Sigma(\hat{\mu}) = \text{diag}(y) \). The natural parameter vector \( \eta \) equals \( \log(\mu) \), so \( \hat{\eta} = \hat{\eta} = \log(y) \), and the function \( \mu \) mapping \( \eta \) to \( \mu \) is \( \mu(\eta) = e^\eta \).

**Notation:** \( \log(\mu) = (\log(\mu_1), \log(\mu_2), \ldots, \log(\mu_p)), \quad e^\eta = (e^{\eta_1}, e^{\eta_2}, \ldots, e^{\eta_p}), \) and in general \( f(x) = (f(x_1), f(x_2), \ldots, f(x_p)) \) for any real-valued function \( f \) of a vector \( x \).

**Multinomial.** Suppose that \( n \) subjects are independently classified into one of \( p \) disjoint categories, with unknown probability \( \pi_j \) for category \( j = 1, 2, \ldots, p \). The sufficient vector is \( y = (y_1, y_2, \ldots, y_p)' \), \( y_j \) being the number of subjects classified into category \( j \); \( y \) has the multinomial distribution on \( p \) categories, \( n \) draws, true probability vector \( \pi \),

\[
y \sim \text{Mult}_p(n, \pi). \tag{3.3}
\]

The mean vector \( \mu \) equals \( n \pi \), so \( \mu_j/n \) is the true probability \( \pi_j \) 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 (3.2). We can use the Poisson specifications in Table 2 to handle the multinomial case. One point of caution is necessary: the function \( tt \) used in the abc program call must be in homogeneous form

\[
tt(cy) = tt(y) \quad \text{for} \quad c > 0. \tag{3.4}
\]

If this is not true, then the homogenized form \( tth \) should be substituted for \( tt \),

\[
tth(y) \equiv tt(ny/\sum_{j=1}^{p} y_j). \tag{3.5}
\]

If \( tt(y) = y_1/n \) for instance, then \( tth(y) = y_1/\sum_{j=1}^{p} y_j \).

As an example, suppose that we observe a \( 2 \times 2 \) contingency table, with the four cells labelled

\[
\begin{array}{cc}
1 & 2 \\
3 & 4
\end{array}
\tag{3.6}
\]

The data vector \( y' = (y_1, y_2, y_3, y_4) \) can be considered either multinomial (3.3), or Poisson (3.2). Figure 4 concerns setting confidence intervals for the cross-product ratio \( \theta = \pi_1 \pi_4/\pi_2 \pi_3 = \mu_1 \mu_4/\mu_2 \mu_3 \).
having observed $y' = (10, 24, 20, 12)$, so $\hat{\theta} = .25$. Here $tt(y) = y_1y_4/y_2y_3$ is already in homogeneous form. The ABC and standard intervals are quite different in this case (but not if the parameter of interest is changed from $\theta$ to $\log \theta$).

This situation allows an exact confidence interval, for $\theta$, conditional on the marginals $y_1 + y_2 = 34, y_1 + y_3 = 30, y_1 + y_2 + y_3 + y_4 = 66$; see Section 4.6 of Lehmann (1986). Given these quantities, $y_1$ has a density depending only on $\theta$, $f_{\theta}(y_1) = e^{(y_1^{(34)})(30-y_1)}y_1^{(32)}$ for $y_1 = 0, 1, 2, \cdots, 30.$ The actual level of a confidence limit $\theta_0$ is defined to be the probability of $y_1$ exceeding the observed value 10 for parameter value $\theta_0$,

$$\text{actual level}(\theta_0) = .5f_{\theta_0}(10) + \sum_{11}^{30} f_{\theta_0}(z), \quad (3.7)$$

where we have split the probability atom at 10. The right side of Figure 4 shows that the actual levels of the ABC limits are very near their nominal values.

![Figure 4](image)

**Figure 4.** Approximate confidence intervals for the cross-product ratio, having observed the multinomial $2 \times 2$ table $(10, 24, 20, 12)$, labelled as in (3.6), $\hat{\theta} = .25.$ The ABC limits, solid curves, have actual levels nearly equal the nominal values; the standard limits are quite inaccurate in this case. Actual levels are calculated according to formula (3.7).

**Binomial.** We observe independent binomial variates,

$$y_j \sim Bi(n_j, \pi_j) \quad \text{independently} \quad j = 1, 2, \cdots, p. \quad (3.8)$$

The data vector $y = (y_1, y_2, \cdots, y_p)'$ has mean vector $\mu = (\mu_1, \mu_2, \cdots, \mu_p)'$ with $\mu_j = n_j \pi_j$. The covariance matrix is $\Sigma(\mu) = \text{diag}(n_j \pi_j(1 - \pi_j)) = \text{diag}(\mu_j(1 - \mu_j/n_j))$, so $S = \text{diag}(y_j(1 - y_j/n_j))$. The natural parameter vector $\eta$ has components $\eta_j = \log(\pi_j/(1 - \pi_j)) = \log(\mu_j/(n_j - \mu_j))$, so $\eta_{\hat{\theta} j} = \log(y_j/(n_j - y_j))$; the mapping $\mu$ from $\eta$ to $\mu$ is $\mu_j = n_j/(1 + e^{\eta_j})$ for $j = 1, 2, \cdots, p.$

**Logistic Regression.** A standard logistic regression analysis links independent binomial observations by a linear regression model for the logistic parameters, see Cox (1970): we observe
independent binomials

\[ z_i \sim \text{Bi}(n_i, \pi_i) \quad \text{independently for} \quad i = 1, 2, \ldots, N, \tag{3.9} \]

and also \( p \times 1 \) covariate vectors \( x_i \), conveniently considered to be the columns of a \( p \times N \) matrix \( X \equiv (x_1, x_2, \ldots, x_N) \). The logit for the \( i \)th case is assumed to be a linear function of \( x_i \) and an unknown regression vector \( \eta \),

\[ \log(\pi_i/(1 - \pi_i)) = x_i'\eta \quad \text{for} \quad i = 1, 2, \ldots, N. \tag{3.10} \]

Situation (3.9) (3.10) is a \( p \)-parameter exponential family with sufficient vector \( y = Xz = \sum_{i=1}^{N} x_i z_i \), and natural parameter vector \( \eta \). A standard logistic regression program provides the MLE \( \hat{\eta} = \text{etahat} \) of \( \eta \). (There is no closed-form expression for \( \hat{\eta} \), but see (3.14) below.) The estimated covariance matrix \( S \) equals \( X \text{diag}(n_i \hat{\pi}_i(1 - \hat{\pi}_i))X' = \sum_i x_i n_i \hat{\pi}_i(1 - \hat{\pi}_i)x_i' \), where \( \hat{\pi}_i \) is the MLE for the \( i \)th component probability

\[ \pi_i = 1/(1 + e^{-z_i'\eta}). \tag{3.11} \]

Most logistic regression programs provide the matrix \( S \) in their outputs. The function \( \mu \) has the simple form

\[ \mu = n/(1 + e^{-z_i'\eta}), \tag{3.12} \]

or equivalently, \( \mu_i = n_i/(1 + e^{x_i'\eta}) \) for \( i = 1, 2, \ldots, N \). With these specifications, \( abc(tt, y, S, \mu, \cdots) \) produces second-order correct confidence intervals for parameters \( \theta = t(\mu) \).

As a simple example, consider the field goal data, Efron (1982): a professional football player, Don Cockcroft, tried 100 field goals over a four season period, with the following record of success from various distances away from the goal,

\[
\begin{array}{cccccc}
\text{attempts/successes} & (x_i/n_i) & 1/4 & 8/27 & 15/32 & 22/25 & 10/12 \\
\text{distance in yards} & (d_i) & 55 & 45 & 35 & 25 & 12 \\
\end{array}
\tag{3.13}
\]

We model the data as in (3.9), (3.10), with \( x_i' = (1, d_i - 30) \). In other words, we assume that Cockcroft’s successes for any given distance were binomially distributed, with probability of success from distance \( d \) given by the logistic formula

\[ \pi(d) = \frac{1}{1 + e^{-[\eta_1 + \eta_2(d - 30)]}}. \tag{3.14} \]

Figure 5 shows the ABC intervals and the standard intervals for two parameters of interest, \( \theta = \eta_2 \), the distance coefficient in (3.13), and \( \theta = \pi(55) \), Cockcroft’s probability of success from 55 yards. The ABC and standard intervals agree for \( \theta = \eta_2 \) (this happened both times we let \( \theta \) equal one component of the natural parameter vector), but disagree considerably for \( \theta = \pi(55) \). The upper limit of the 95% interval for \( \pi(55) \) is .319 for ABC, versus .266 for the standard.

The parameters \( \eta_2 \) and \( \pi(55) \) are expressed as functions of the natural vector \( \eta \), rather than of the expectation vector \( \mu \), say \( s(\eta) \) instead of \( t(\mu) \). In order to use \( abc(tt, y, S, \mu, \cdots) \) we need to evaluate the implied function \( tt(y + \Delta) = s(\hat{\eta}(y + \Delta)) \) for various choices of the incremental vector.
\( \Delta \), where \( \hat{\eta}(y + \Delta) \) is the MLE of \( \eta \) corresponding to sufficient statistic \( y + \Delta \). Since \( \Delta \) 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
\[
\hat{\eta}^{(2)} = \hat{\eta}^{(1)} + S^{-1}(y + \Delta - m u(\hat{\eta}^{(1)})).
\]  
(3.15)

The computations giving Figure 5 did four iterations of (3.15) for each choice of \( \Delta \), beginning with \( \hat{\eta}^{(0)} = \hat{\eta} \), see the program "etalogit" in the appendix.

![Diagram with distance coefficient and probability success](image)

**Figure 5.** ABC and standard intervals for two parameters, linear logistic regression (3.13) for field-goal data (3.12): \( \theta = \eta_2 \), the distance coefficient, left panel; and \( \theta = \pi(55) \), the probability of success from \( d = 55 \) yards, right panel. The ABC and standard intervals agree in the first case, but not in the second.

Four iterations of (3.14) is plenty for the small values of \( \Delta \) needed to evaluate the numerical derivatives in the abc algorithm. However, four may not be enough for the final evaluation of the ABC limits, (2.21), since the increment \( \Delta = (\hat{\delta}/\hat{\sigma})z_{1\alpha} \) can then be large. The abc program of the appendix can be modified to use more iterations of (3.14) for the final evaluation (2.21). An alternative is to use the quadratic approximation (2.26) in place of \( t(\hat{\mu} + (\hat{\delta}/\hat{\sigma})z_{1\alpha}) \). The abc program includes the quadratic values (2.26) in the output.

**Multivariate Normal.** Suppose that the data consists of \( n \) independent observations from a \( d \)-dimensional multivariate normal distribution with unknown mean vector \( \lambda \) and unknown covariance matrix \( \Gamma \),
\[
x_i \sim N_d(\lambda, \Gamma) \quad \text{independently for } \quad i = 1, 2, \cdots, n.
\]  
(3.16)

(so \( d = 1 \) is the univariate normal case.) The sufficient vector \( y \) has \( p = d(d + 3)/2 \) coordinates, say \( y^\prime = (y_1^\prime, y_2^\prime) \) where \( y_1 \) is the \( d \)-dimensional vector of means and \( y_2 \) is the \( d(d + 1)/2 \) dimensional vector of mean squares.
\[
y_1 = \frac{1}{n} \sum_{i=1}^{n} x_i/n = \bar{x} \quad \text{and} \quad y_2 = \left( \sum_{i=1}^{n} x_i x_i^\prime/n \right)^{(u)}.
\]  
(3.17)
Here \( a = A^{(v)} \) indicates the \( d(d+1)/2 \) dimensional vector obtained by ordering the super-diagonal elements of the symmetric \( d \times d \) matrix \( A \) in any convenient way, say

\[
a' = (A_{11}, A_{12}, A_{13}, \ldots, A_{22}, A_{23}, \ldots, A_{dd});
\]  

(3.17a)

The inverse mapping for \( a \) back to \( A \) will be denoted by \( A = a^{(m)} \).

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

\[
\mu_1 = \lambda \quad \text{and} \quad \mu_2 = (\lambda \lambda' + \Gamma)^{(v)}.
\]

(3.18)

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

\[
\eta_1 = n \Gamma' \lambda \quad \text{and} \quad \eta_2 = n \{ \text{diag}(\Gamma^{-1})/2 - \Gamma^{-1} \}^{(v)},
\]

(3.19)

where \( \text{diag}(A) \) indicates the diagonal matrix obtained from a square matrix \( A \) by setting its off-diagonal elements equal to zero. The mapping \( m \mu \) from \( \eta \) to \( \mu \) is expressed in two steps, first from \( \eta \) to \( (\lambda, \Gamma) \),

\[
\Gamma = -n \{ \text{diag}(\eta_2^{(m)} + \eta_2^{(m)}) \} \quad \text{and} \quad \lambda = \frac{1}{n} \Gamma(\eta_1),
\]

(3.20)

and then from \( (\lambda, \Gamma) \) to \( \mu \) by (3.17). Likewise the mapping for \( \mu \) to \( \eta \), or from \( \tilde{\mu} \) to \( \tilde{\eta} \), is given by

\[
\lambda = \mu_1 \quad \text{and} \quad \Gamma^{-1} = \{ \mu_2^{(m)} - \mu_1 \mu_1' \}^{-1},
\]

(3.21)

followed by (3.19).

In this case, and in others involving the more elaborate exponential families, it is easiest to calculate the covariance matrix \( \Sigma \) and the matrix \( S = \Sigma(\hat{\mu}) \) needed in the abc program by differentiation of \( \mu(\eta) \),

\[
\Sigma_{\eta x \eta} = \hat{\mu}(\eta) = (\partial \mu_i / \partial \eta_j),
\]

(3.22)

(2.6). The program Sdiff given in the appendix calculates \( S \) by the numerical differentiation of \( m \mu(\eta) \) at \( \eta = \eta \).

Alternatively, \( \Sigma \) (and \( S \)) for the multivariate normal can be calculated from lengthy but straightforward formulas. Define \( \xi_{ij} = \Gamma_{ij}/\lambda_i \lambda_j \) for \( i, j = 1, 2, \ldots, d \). Then

\[
\Sigma \equiv \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}
\]

(3.23)

where the four submatrices are defined as follows. The \( d \times d \) matrix \( \Sigma_{11} \) is

\[
\Sigma_{11} = \Gamma/n.
\]

(3.23a)

The \( d \times d(d+1)/2 \) matrix \( \Sigma_{12} \) has elements

\[
\Sigma_{12,ij} = \lambda_i \lambda_j \lambda_k (\xi_{ik} + \xi_{ij})/n,
\]

(3.23b)
where \( i = 1, 2, \ldots, d \) and \( jk = (11, 12, 13, \ldots, 1d, 22, 23, \ldots, dd) \), as in (3.16a); \( \Sigma 21 = \Sigma 12' \); and the \( d(d+1)/2 \times d(d+1)/2 \) matrix \( \Sigma 22 \) has elements

\[
\Sigma 22_{ij,k\ell} = \lambda_i \lambda_j \lambda_k \lambda_{\ell} [(1 + \xi_{ik})(1 + \xi_{j\ell}) + (1 + \xi_{ik})(1 + \xi_{j\ell}) - 2]/n, \tag{3.23c}
\]

for \( ij \) and \( k\ell \) equalling \((11, 12, 13, \ldots, dd)\). See Section 7.3 of Anderson (1958).

As an example, we consider the score data from Mardia, Kent, and Bibby (1979). The data consists of \( d = 5 \) test scores, for each of \( n = 88 \) students. Starting with the multivariate normal model (3.15), we consider two parameters of interest, \( \theta_1 = \) maximum eigenvalue of \( \Gamma \) and \( \theta_2 = \theta_1 / tr(\Gamma) \), (so \( \theta_2 = \) max eigenvalue of \( \Gamma \)/sum of eigenvalues of \( \Gamma \), the proportion of the total variance attributable to the first principal direction). Figure 6 compares the ABC intervals with the standard intervals. For \( \theta_1 \) the ABC intervals are shifted noticeably up from the standard, while \( \theta_2 \) shows a smaller downward shift. In both cases, but particularly for \( \theta_2 \), even the standard intervals are not usually available to the statistician.

Note: \( \hat{\Gamma} \), the MLE of \( \Gamma \), equals \( \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})' / n \) rather than the unbiased estimate \( \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z}) / (n - 1) \). The ABC algorithm automatically includes this downward bias in the calculation of the bias-correction constant \( z_0 \). (This partially accounts for the accuracy of the ABC intervals for \( \text{var} \{\text{LSAT} \} \) seen in the right panel of Figure 1.)

**Figure 6.** Approximate confidence intervals for two parameters in the multivariate normal model \( x_i \sim i.i.d. N_d(\lambda, \Gamma), i = 1, 2, \ldots, n; \theta_1 = \max \text{eigenvalue of } \Gamma \), left panel. and \( \theta_2 = \theta_1 / tr(\Gamma) \), right panel. Data is the score data, Mardia, Kent, and Bibby (1979), with \( n = 88 \) and \( d = 5 \).

**Gamma.** Suppose we observe independent scaled gamma variates, each with its own unknown scale parameter.

\[
y_i \sim \mu_i G_{n_i} / n_i \quad \text{independently} \quad i = 1, 2, \ldots, p. \tag{3.24}
\]

17
(Each $y_i$ might be an average $\bar{x} = \sum_{j=1}^{m_i} x_{ij}/m_i$ where $x_{ij} \sim \mathcal{N}_{i.i.d.}(\mu_i, \sigma_i)$ some known shape parameter, as in (2.10), in which case $n_i = m_i \mu_i$.) Then the call to program abc uses $y = (y_1, y_2, \ldots, y_p)'$, $S = \text{diag}(y_i^2/n_i)$, $\eta = -n_j/y_j$ for $j = 1, 2, \ldots, p$, and $\mu(\eta)_j = -n_j/\eta_j$ for $j = 1, 2, \ldots, p$.

4. Theoretical Details.

Consider a parametric model indexed by a vector $\eta = (\eta_1, \ldots, \eta_p)'$, and suppose that the log-likelihood function for $\eta$ based on an observed random variable $y = (y_1, \ldots, y_p)'$ has the exponential family form

$$\ell(\eta; y) = n\{\eta'y - \tilde{\psi}(\eta)\}, \quad (4.1)$$

where $\eta$ and $\tilde{\psi}(\eta)$ are both $O(1)$ and $y$ is $O_p(n^{-1/2})$. If we define $\eta = n\bar{\eta}$ and $\psi(\eta) = n\tilde{\psi}(\eta) = n\tilde{\psi}(\eta)/n$, then the log-likelihood function for $\eta$ from (4.1) is

$$\ell(\eta; y) = \eta'y - \psi(\eta), \quad (4.2)$$

which agrees with (2.4). The cumulant generating function of $y$ is

$$\log E(e^{t'y}) = \psi(\eta + \xi) - \psi(\eta), \quad \xi = (\xi_1, \ldots, \xi_p)', \quad (4.3)$$

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

$$E(y_i) = \psi_i(\eta) = \tilde{\psi}_i(\eta), \quad \text{var}(y_i, y_j) = \psi_{ij}(\eta) = \tilde{\psi}_{ij}(\eta)/n, \quad (4.4)$$

$$\text{cum}(y_i, y_j, y_k) = \psi_{ijk}(\eta) = \tilde{\psi}_{ijk}(\eta)/n^2, \quad (i, j, k = 1, \ldots, p),$$

where $\psi_i(\eta) = \partial \psi(\eta)/\partial \eta_i$, $\tilde{\psi}_i(\eta) = \partial \tilde{\psi}(\eta)/\partial \eta_i$, $\psi_{ij}(\eta) = \partial^2 \psi(\eta)/\partial \eta_i \partial \eta_j$, $\tilde{\psi}_{ij}(\eta) = \partial^2 \tilde{\psi}(\eta)/\partial \eta_i \partial \eta_j$, etc. Fourth- and higher-order joint cumulants are $O(n^{-3})$ or smaller. It is convenient to work with the expectation parameter $\mu = (\mu_1, \ldots, u_p)' = E(y)$, whose components are defined by $\mu_i = \psi_i(\eta)$ ($i = 1, \ldots, p$). The maximum likelihood equations $\hat{y}_i = \psi_i(\hat{\eta})$ ($i = 1, \ldots, p$) obtained from (4.2) yield $\hat{\mu} = y$. The notation given here coincides with that introduced in formulas (2.5)-(2.7).

Now suppose that interest lies in constructing confidence limits for a scalar parameter $\theta = t(\mu)$. Let $t_i(\mu) = \partial t(\mu)/\partial \mu_i$ and $t_{ij}(\mu) = \partial^2 t(\mu)/\partial \mu_i \partial \mu_j$ ($i, j = 1, \ldots, p$). By adopting the common notational convention in which summation is assumed over the range $1, \ldots, p$ for repeated indices, the Cramer-Rao lower bound (2.8) can be written $\sigma = \{\psi_{ij}t ihtk}\{t\}_{ij}^{1/2}$. To facilitate notation, this convention is used freely throughout the present section. Now $\sigma$ is $O(n^{-1/2})$, and the variance of the maximum likelihood estimator $\hat{\theta} = t(\hat{\mu})$ is $\sigma^2 + O(n^{-2})$. Cumulants (4.4) can be used in standard calculations to show that, with error of order $O(n^{-1})$, the first three cumulants of $U = (\hat{\theta} - \theta)/\sigma$ are

$$E(U) = \frac{1}{2}m_2, \quad \text{var}(U) = 1, \quad \text{skew}(U) = m_1 + 3m_3, \quad (4.4)$$

and those of $V = (\hat{\theta} - \theta)/\sigma$ are

$$E(V) = -\frac{1}{2}m_1 + \frac{1}{2}m_2 - m_3, \quad \text{var}(V) = 1, \quad \text{skew}(V) = -2m_1 - 3m_3, \quad (4.5)$$

where

$$m_1 = \psi_{ijk}t_{ij}tkt_k/\sigma^3, \quad m_2 = \psi_{ij}t_{ij}/\sigma, \quad m_3 = \psi_{ik}\psi_{jkt}t_{ikt}/\sigma^3. \quad (4.6)$$
Note that $m_1, m_2,$ and $m_3$ are all $O(n^{-1/2})$. Fourth- and higher-order cumulants of $U$ and $V$ are $O(n^{-1})$ or smaller.

If $\hat{\theta}$ has a continuous distribution, then exact confidence limits for $\theta$ can be based on the distribution of $V = (\hat{\theta} - \theta)/\hat{\sigma}$. Let $K_\eta$ be the distribution function of $V$, that is, $K_\eta(x) = \text{Prob}_\eta(V \leq x)$. The quantity

$$\hat{\theta}_{EX}[1 - \alpha] = \hat{\theta} - \hat{\sigma} K_\eta^{-1}(\alpha)$$

(4.7)

is an exact upper $1 - \alpha$ confidence limit for $\theta$, since $\text{Prob}_\eta\{K_\eta^{-1}(\alpha) \leq (\hat{\theta} - \theta)/\hat{\sigma}\} = 1 - \alpha$ implies $\text{Prob}_\eta\{\theta \leq \hat{\theta}_{EX}[1 - \alpha]\} = 1 - \alpha$. For example (2.11), with $\theta = \mu$, we have that $V = n^{1/2}\{1 - (\mu/y)\}$ is pivotal, so $K_\mu(x) = \text{Prob}\{\chi^2_{(2n)} \leq 2n(1 - n^{-1/2}x)^{-1}\}$ does not depend on $\mu$ and $\hat{\mu}_{EX}[1 - \alpha] = 2n y/\chi^2_{(2n)}$, where $\chi^2_{(2n)}$ is the $\alpha$ quantile of the chi-squared distribution having $2n$ degrees of freedom. For most applications, however, the theoretical exact limit (4.7) cannot be explicitly calculated because it depends on the unknown parameter $\eta$ through $K_\eta^{-1}(\alpha)$. Nonetheless, the notion of second-order correctness for approximate confidence limits can be defined in terms of $\hat{\theta}_{EX}[1 - \alpha]$. This choice of exact confidence limit follows Hall (1988).

In the case that $\hat{\theta}$ is continuously distributed, it follows from (4.5) that $K_\eta^{-1}(\alpha)$ has the Cornish-Fisher expansion

$$K_\eta^{-1}(\alpha) = -\{z_\alpha + \left(\frac{1}{6} m_1 - \frac{1}{2} m_2 + \frac{1}{2} m_3\right) + \left(\frac{1}{3} m_1 + \frac{1}{2} m_3\right) z_\alpha^2\} + O(n^{-1})$$

(4.8)

$$= -\{z_\alpha + z_0 + (2a + \frac{1}{2} m_3) z_\alpha^2\} + O(n^{-1}),$$

where

$$z_0 = \frac{1}{6} m_1 - \frac{1}{2} m_2 + \frac{1}{2} m_3, \quad a = \frac{1}{6} m_1,$$

and, as for (1.1), $z_\alpha$ is the $1 - \alpha$ quantile of the standard normal distribution. Thus

$$\hat{\theta}_{EX}[1 - \alpha] = \hat{\theta} + \hat{\sigma}\{z_\alpha + z_0 + (2a + \frac{1}{2} m_3) z_\alpha^2\} + O_p(n^{-3/2}),$$

(4.9)

since $\hat{\sigma}$ is $O_p(n^{-1/2})$. If we define

$$\text{curv} = \frac{1}{2}(m_2 - m_3),$$

then $z_0 = a - \text{curv}$; moreover, a straightforward calculation shows that

$$z_0 = \Phi^{-1}\{2\Phi(a)\Phi(-\text{curv})\} + O(n^{-1}).$$

(4.10)

When $p = 1$, $z_0 = a$ because $m_2 = m_3$ and curv = 0. Similarly, $z_0 = a$ when $\theta$ is a linear combination of $\mu_1, \cdots, \mu_p$ as $m_2$ and $m_3$ are both 0 in that case.

An approximate upper $1 - \alpha$ confidence limit $\hat{\theta}_A[1 - \alpha]$ is called second-order correct if it differs from $\hat{\theta}_{EX}[1 - \alpha]$ by terms of order $O_p(n^{-3/2})$, or equivalently, of order $O_p(\hat{\sigma}/n)$. The Cornish-Fisher expansion of $K_\eta$ can be used to show that the coverage error of such a limit is $O(n^{-1})$, that is,

$$\text{Prob}_\eta\{\theta \leq \hat{\theta}_A[1 - \alpha]\} = 1 - \alpha + O(n^{-1}).$$
A variety of second-order correct confidence limits is available. One simple method for constructing them is to replace the unknown quantity \( m_i \) in (4.9) by an estimator \( \hat{m}_i \) that satisfies \( \hat{m}_i = m_i + O_p(n^{-1}) \) \((i = 1, 2, 3)\), thereby producing the approximate limit

\[
\hat{\theta} + \hat{\sigma} \{ z_\alpha + \hat{z}_0 + (2\hat{a} + \frac{1}{2} \hat{m}_3)z_\alpha^2 \}.
\]  

(4.11)

The quantities \( \hat{m}_i \) \((i = 1, 2, 3)\) obtained by substituting \( \hat{\eta} \) for the unknown parameter \( \eta \) in (4.6) are obvious candidates for such estimators. Other second-order correct limits are given by the bootstrap-\( t \) and \( BC_\alpha \) methods.

For the bootstrap-\( t \) limit, the parametric bootstrap distribution \( K_\hat{\eta} \) of the “studentized” quantity \( V \) is used to approximate \( K_\eta \) in (4.7). The limit is therefore defined as

\[
\hat{\theta}_{B-t}[1 - \alpha] = \hat{\theta} - \hat{\sigma} K_\hat{\eta}^{-1}(\alpha),
\]

and it follows immediately from (4.8) that \( \hat{\theta}_{B-t}[1 - \alpha] \) is second-order correct. In example (2.11), \( \hat{\theta}_{B-t}[1 - \alpha] = \hat{\theta}_{EX}[1 - \alpha] \), since \( V \) is pivotal for that problem.

The \( BC_\alpha \) limit (2.2)-(2.3) is defined by

\[
\hat{\theta}_{BC_\alpha}[1 - \alpha] = G_\hat{\eta}^{-1}\{ \Phi(\hat{z}_0 + z_\alpha) \}, \quad z_\alpha = \frac{\hat{z}_0 + z_\alpha}{1 - \hat{a}(\hat{z}_0 + z_\alpha)},
\]

(4.12)

where \( G_\eta \) is the distribution function of \( \hat{\theta} \), that is, \( G_\eta(\eta) = \text{Prob}_\eta(\hat{\theta} \leq \eta) \), and \( G_\hat{\eta} \) is the parametric bootstrap distribution function. In (4.12), \( \hat{z}_0 \) and \( \hat{a} \) are estimators that differ from \( z_0 \) and \( a \) respectively by terms of order \( O_p(n^{-1}) \). Now let \( H_\eta \) be the distribution function of \( U = (\hat{\theta} - \theta)/\sigma \), so that \( G_\eta^{-1}(\alpha) = \theta + \sigma H_\eta^{-1}(\alpha) \). By again assuming \( \hat{\theta} \) to be continuous, it follows from (4.4) that \( H_\eta^{-1}(\alpha) \) has the Cornish-Fisher expansion

\[
H_\eta^{-1}(\alpha) = -\{ z_\alpha + (\frac{1}{6} m_1 - \frac{1}{2} z_0 + \frac{1}{2} z_\alpha) - (\frac{1}{6} m_1 + \frac{1}{2} m_3)z_\alpha^2 \} + O(n^{-1})
\]

\[
= -\{ z_\alpha + z_0 - (a + \frac{1}{2} m_3)z_\alpha^2 \} + O(n^{-1}).
\]

Since

\[
\hat{z}_0 + z_\alpha = z_\alpha + 2\hat{z}_0 + \hat{a}z_\alpha^2 + O_p(n^{-1}),
\]

we have

\[
\hat{\theta}_{BC_\alpha}[1 - \alpha] = \hat{\theta} + \hat{\sigma} H_\hat{\eta}^{-1}\{ \Phi(\hat{z}_0 + z_\alpha) \}
\]

\[
= \hat{\theta} + \hat{\sigma} \{ z_\alpha + 2\hat{z}_0 + az_\alpha^2 - \hat{z}_0 + (\hat{a} + \frac{1}{2} \hat{m}_3)z_\alpha^2 \} + O_p(n^{-3/2})
\]

\[
= \hat{\theta} + \hat{\sigma} \{ z_\alpha + z_0 + (2a + \frac{1}{2} m_3)z_\alpha^2 \} + O_p(n^{-3/2})
\]

\[
= \hat{\theta}_{EX}[1 - \alpha] + O_p(n^{-3/2}).
\]

Therefore, \( \hat{\theta}_{BC_\alpha}[1 - \alpha] \) is second-order correct.

Second-order correctness of bootstrap-\( t \) and \( BC_\alpha \) confidence limits is discussed in parametric and nonparametric settings by DiCiccio and Romano (1988) and Hall (1988).

20
Unfortunately, direct calculation of the bootstrap-t and \( BC_\alpha \) limits is made difficult by the fact that the bootstrap distributions \( K_\eta \) and \( G_\eta \) do not in general have closed form expressions. To determine these bootstrap distributions by simulation is cumbersome. Similarly, calculation of the estimates \( \hat{m}_i \) \((i = 1, 2, 3)\) for use in an approximate limit such as (4.11) by setting \( \eta = \hat{\eta} \) in (4.6) is tedious, since it involves third- and fourth-order sums of second- and third-order partial derivatives. For the ABC method, these estimators are expressed in terms of simple derivatives that can be computed numerically.

The derivatives used in the ABC method are now presented. By definition,

\[
\hat{\psi}_i = \frac{dt(\hat{\eta} + \lambda c^i)h/d\lambda|_{\lambda=0}}{(i = 1, \cdots, p),}
\]

where \( c_i \) is the unit vector with \( i \)th component 1. The relationship \( \mu_i = \psi_i(\eta) \) implies

\[
\hat{\psi}_{ij} = \frac{d\mu_i(\hat{\eta} + \lambda c^j)h/d\lambda|_{\lambda=0}}{(i, j = 1, \cdots, p)}.
\]

As in Section 2, let \( \hat{\delta}_{i} = \hat{\psi}_{ij} \hat{t}_j \) \((i = 1, \cdots, p)\), and put \( \hat{\epsilon} = \hat{t}_1, \cdots, \hat{t}_p \)' and \( \delta = (\delta_1, \cdots, \delta_p)' \), and \( \hat{\Sigma} = (\hat{\psi}_{ij}) \), so that \( \hat{\Sigma} \) is a \( p \times p \) matrix. Then

\[
\hat{\sigma}^2 = \hat{\psi}_{i} \hat{t}_i \hat{\epsilon}_j = d^2 \psi(\hat{\eta} + \lambda \hat{\epsilon})h/d\lambda^2|_{\lambda=0},
\]

\[
\hat{m}_1 = \hat{\psi}_{i} \hat{t}_i \hat{\epsilon}_j \hat{\delta}^3 = \{d^2 \mu_i(\hat{\eta} + \lambda \hat{\epsilon})h/d\lambda^2|_{\lambda=0}\}/\hat{\sigma}^3.
\]

Take \( D \) to be a diagonal matrix of eigenvalues of \( \hat{\Sigma} \) and \( \Gamma \) to be an orthogonal matrix whose columns consist of corresponding eigenvectors. Therefore, \( \hat{\Sigma} = \Gamma D \Gamma' = CC' \), where \( C = \Gamma D^{1/2} \). Let \( C_1, \cdots, C_p \) be the columns of \( C \). Then

\[
\hat{\psi}_{ij} = \sum_{k=1}^{p} C_i C_j^k (i, j = 1, \cdots, p),
\]

and

\[
\hat{m}_2 = \hat{\psi}_{ij} \hat{t}_i \hat{\epsilon}_j / \hat{\sigma} = \{\sum_{k=1}^{p} d^2 \mu_i(\hat{\eta} + \lambda C^k)h/d\lambda^2|_{\lambda=0}\}/\hat{\sigma}.
\]

Finally,

\[
\hat{m}_3 = \hat{\psi}_{i} \hat{t}_i \hat{\epsilon}_j \hat{\delta}^3 = \{d^2 \mu_i(\hat{\eta} + \lambda \hat{\epsilon})h/d\lambda^2|_{\lambda=0}\}/\hat{\sigma}^3.
\]

The derivatives given in (4.13)-(4.18) can easily be computed by numerical differentiation, which is an essential aspect of the ABC algorithm. Having computed \( \hat{m}_i \) \((i = 1, 2, 3)\) in this way, we then obtain

\[
\hat{a} = \frac{1}{\hat{m}_1}, \quad \text{curv}(\hat{\mu}) = \frac{1}{2}(\hat{m}_2 - \hat{m}_3), \quad \hat{z}_0 = \hat{a} - \text{curv}(\hat{\mu}),
\]

as described in (2.30), (2.31), and (2.33), respectively. Another estimate of \( z_0 \) is \( \hat{z}_0 \) \( = \Phi^{-1}[2\Phi(\hat{a})\Phi\{-\text{curv}(\hat{\mu})\}] \), which appears at (2.32). This estimate, like \( \hat{z}_0 \), differs from \( z_0 \) by \( O_p(n^{-1}) \), by virtue of (4.10).

The ABC limit (2.21) is

\[
\hat{\theta}_{ABC}[1 - \alpha] = \frac{\hat{\theta} + \hat{\epsilon} / \hat{\sigma}}{z_{1\alpha}}
\]

where \( z_{1\alpha} = z_0 + \alpha z_{0\alpha} \) satisfies

\[
z_{1\alpha} = z_0 + 2az_{\alpha}^2 + O(n^{-1}).
\]
Therefore,
\[
\hat{\theta}_{ABC}[1 - \alpha] = t(\hat{\mu}) + \frac{i_i \hat{\delta}_i}{\hat{\sigma}} z_1 \alpha + \frac{1}{2} \frac{i_i \hat{\delta}_i \hat{\delta}_j}{\hat{\sigma}^2} z_1^2 \alpha + O_p(n^{-3/2})
\]
\[
= \hat{\theta} + \hat{\sigma} \{ z_\alpha + z_0 + (2a + \frac{1}{2}m_3) \hat{z}_\alpha^2 \} + O_p(n^{-3/2}),
\]
since \(i_i \hat{\delta}_i / \hat{\sigma} = \hat{\sigma}\) and \(i_i \hat{\delta}_i \hat{\delta}_j / \hat{\sigma}^3 = m_3 + O(n^{-1})\). Comparison of (4.19) with (4.9) shows that \(\hat{\theta}_{ABC}[1 - \alpha]\) is second-order correct.

It can be argued that choosing to define second-order correctness in terms of the theoretical exact limit \(\hat{\theta}_{EX}[1 - \alpha]\) is somewhat arbitrary. We now show, however, that some other commonly used approximate confidence limits are second-order correct according to this definition, and thus the bootstrap-t, \(BC_a\) and ABC limits all agree with these other ones to second order.

An especially useful method for constructing confidence limits is based on a normal approximation to the distribution of the signed root of the likelihood ratio statistic. The profile likelihood for the parameter \(\theta\) is \(\bar{\ell}(\theta) = \ell(\hat{\eta}; y)\), where \(\hat{\eta} = \hat{\eta}(\theta)\) maximizes the log likelihood function \(\ell(\eta; y)\) subject to the constraint \(t(\mu) = \theta\). The likelihood ratio statistic for \(\theta\) is
\[
W = 2\{\ell(\hat{\eta}; y) - \ell(\bar{\eta}; y)\} = 2\{\bar{\ell}(\hat{\theta}) - \bar{\ell}(\theta)\},
\]
and its signed root is given by
\[
R = \text{sgn}(\hat{\theta} - \theta) W^{1/2}.
\]
A lengthy calculation shows that the first three derivatives of \(\bar{\ell}(\theta)\) evaluated at \(\hat{\theta}\) are
\[
\bar{\ell}^{(1)}(\hat{\theta}) = 0, \quad \bar{\ell}^{(2)}(\hat{\theta}) = -1/\hat{\sigma}^2, \quad \bar{\ell}^{(3)}(\hat{\theta}) = (2\hat{m}_1 + 3\hat{m}_3)/\hat{\sigma}^3,
\]
and hence \(R\) has the expansion
\[
R = V + \left(\frac{1}{3}m_1 + \frac{1}{2}m_3\right)V^2 + O_p(n^{-1}).
\]  
(4.20)
The first three cumulants of \(R\) are easily found from (4.5) and (4.20) to be
\[
E(R) = -z_0 + O(n^{-1}), \quad \text{var}(R) = 1 + O(n^{-1}), \quad \text{skew}(R) = O(n^{-1}),
\]
while the fourth- and higher-order cumulants of \(R\) are \(O(n^{-1})\) or smaller. Thus the standard normal approximation to the distribution of \(R + z_0\) has error of order \(O(n^{-1})\).

The likelihood-ratio approximate upper \(1 - \alpha\) confidence limit \(\hat{\theta}_{LR}[1 - \alpha]\) is that value of \(\theta\) for which \(R + z_0 = -z_\alpha\). It is clear from the accuracy of the standard normal approximation that the coverage error of \(\hat{\theta}_{LR}[1 - \alpha]\) is \(O(n^{-1})\), that is, \(\hat{\theta}_{LR}[1 - \alpha]\) is second-order accurate. Moreover, solving
\[
-z_\alpha = V + \left(\frac{1}{3}m_1 + \frac{1}{2}m_3\right)V^2 + O_p(n^{-1})
\]
for \(\theta\) yields
\[
\hat{\theta}_{LR}[1 - \alpha] = \hat{\theta} + \hat{\sigma} \{ z_\alpha + (\frac{1}{3}m_1 + \frac{1}{2}m_3) \hat{z}_\alpha^2 \} + O_p(n^{3/2}),
\]
and therefore \(\hat{\theta}_{LR}[1 - \alpha]\) is second order correct.
Mean adjustments to the signed roots of likelihood ratio statistics have been discussed in general by Sprott (1973), DiCiccio (1984), McCullagh (1984), and Barndorff-Nielsen (1986).

The parameter $\hat{\eta}$ in (4.1) is often called the canonical parameter. Suppose that a component of $\hat{\eta}$, say $\theta = \eta_1$, is the parameter of interest and $\lambda = (\eta_2, \ldots, \eta_p)$ is a nuisance parameter. In this case

$$
\sigma^2 = n^{-2} \psi_{11}, \quad m_1 = n^{-3} \psi_{i j k} \psi_{j 1} \psi_{k 1}, \quad m_2 = -n^{-2} \psi_{i j k} \psi_{i j} \psi_{k 1} / \sigma, \quad m_3 = -m_1,
$$

$$
z_0 = -\frac{1}{3} m_1 - \frac{1}{2} m_2, \quad a = \frac{1}{6} m_1, \quad \text{curv} = \frac{1}{2} (m_1 + m_2),
$$

where $\Sigma^{-1} = (\psi_{i j})^{-1} = (\psi_{i j}^T)$.

The conditional distribution of $y_1$ given $y_2, \ldots, y_p$ does not depend on the nuisance parameter $\lambda$, and hence it can be used for inference about $\eta_1$. Barndorff-Nielsen and Cox (1979) have used a double saddlepoint method to approximate the conditional log likelihood function for $\theta = \eta_1$. Their approximation is

$$
\tilde{\ell}(\theta) = \frac{1}{2} \log \det \Sigma_{\lambda \lambda}(\hat{\eta}(\theta)) + \tilde{\ell}(\theta),
$$

where $\Sigma_{\lambda \lambda}$ is the $(p-1) \times (p-1)$ submatrix of $\Sigma$ corresponding to $\lambda$, $\hat{\eta}$ maximizes the log likelihood function $\ell(\eta; y)$ for fixed $\theta = \eta_1$, and $\tilde{\ell}(\theta)$ is the profile likelihood function for $\theta = \eta_1$. Up to an additive constant, the conditional log likelihood function has the expansion

$$
-\frac{1}{2} V^2 + \frac{1}{2} (m_1 + m_2) V + \frac{1}{6} m_1 V^3 + O_p(n^{-1}), \quad (4.21)
$$

and the conditional maximum likelihood estimate of $\theta$ given $y_2, \ldots, y_p$ is

$$
\bar{\theta} = \hat{\theta} - \frac{1}{2} \hat{\sigma} (m_1 + m_2) + O_p(n^{-3/2}). \quad (4.22)
$$

The conditional likelihood ratio statistic for $\theta$ is

$$
\bar{W} = 2\{\tilde{\ell}(\theta) - \tilde{\ell}(\theta)\}
$$

and its signed square root is

$$
\bar{R} = \text{sgn}(\bar{\theta} - \theta) \bar{W}^{1/2}.
$$

Now (4.21) and (4.22) can be used to show that

$$
\bar{R} = V - \frac{1}{2} (m_1 + m_2) - \frac{1}{6} m_1 V^2 + O_p(n^{-1}),
$$

whereas, from (4.20),

$$
R = V - \frac{1}{6} m_1 V^2 + O_p(n^{-1}).
$$

Note that $\bar{R} = R - \frac{1}{2} (m_1 + m_2) + O_p(n^{-1}) = R - \text{curv} + O_p(n^{-1})$, so to the order of error considered the effect of using $\bar{R}$ is to correct $R$ for the term curv in its expectation $E(R) = -a + \text{curv} + O(n^{-1})$. It follows then that the first three cumulants of $\bar{R}$ are

$$
E(\bar{R}) = -a + O(n^{-1}), \quad \text{var}(\bar{R}) = 1 + O(n^{-1}), \quad \text{skew}(\bar{R}) = O(n^{-1});
$$
fourth- and higher-order cumulants of \( \tilde{R} \) are \( O(n^{-1}) \) or smaller. Therefore, the standard normal approximation to the distribution of \( \tilde{R} + \hat{a} \) has error of order \( O(n^{-1}) \), and this approximation can be used to construct second-order accurate confidence limits for \( \theta = \bar{\eta}_1 \). Since \( \tilde{R} + \hat{a} = R + \bar{z}_0 + O_p(n^{-1}) \), the approximate upper \( 1 - \alpha \) confidence limit for \( \theta \) obtained in this way agrees with \( \hat{\theta}_{LR}[1 - \alpha] \) to second order and hence is second-order correct. It follows from DiCiccio and Martin (1990) that the coverage error of both these approximate limits is \( O(n^{-1}) \) conditionally given \( y_2, \ldots, y_p \) as well as unconditionally. Thus, the ABC method is also conditionally second-order accurate for this problem.

Cox and Reid (1987) have developed a conditional profile likelihood for inference about a scalar parameter in general models when nuisance parameters are present. Their approach requires that the parameter of interest be orthogonal to the nuisance parameters in the sense of expected Fisher information, and the effect of conditioning is to modify the profile likelihood function for the parameter of interest. In the exponential family context, the Cox and Reid conditional profile likelihood for \( \theta = \bar{\eta}_1 \) is the same as the Barndorff-Nielsen and Cox approximate conditional likelihood. However, the conditional profile likelihood is available for other parameters of interest. As for conditional likelihood, the standard normal approximation to the mean adjusted signed root of the conditional profile likelihood ratio statistic produces confidence limits that are second-order correct. We demonstrate this result in the case where a component of the mean vector \( \mu \), say \( \theta = \mu_1 \), is of interest.

If \( \theta = \mu_1 \), then

\[
\sigma^2 = \psi_{11}, \quad m_1 = \psi_{111}/\sigma^2, \quad m_2 = m_3 = 0, \quad z_0 = a = \frac{1}{6} m_1, \quad \text{curv} = 0.
\]

Let

\[
m_4 = \psi_{14}/\sigma.
\]

The parameter \( \lambda = (\bar{\eta}_2, \ldots, \bar{\eta}_p) \) is orthogonal to \( \theta \), and the conditional profile likelihood function for \( \theta \) is

\[
\ell(\theta) = \frac{1}{2} \log \det \Sigma_{\lambda_1 \lambda_1}(\hat{\gamma}(\theta)) + \tilde{\ell}(\theta),
\]

where \( \hat{\gamma}(\theta) \) maximizes \( \ell(\eta; y) \) for fixed \( \theta = \mu_1 \) and \( \tilde{\ell}(\theta) \) is the profile likelihood for \( \theta \). Up to an additive constant, the conditional profile likelihood has the expansion

\[
-\frac{1}{2} V^2 + \frac{1}{2} (-m_1 + m_4) V - \frac{1}{3} m_1 V^3 + O_p(n^{-1}),
\]

which has its maximum value at \( \theta \) that satisfies

\[
\bar{\theta} = \hat{\theta} + \sigma \frac{1}{2} (m_1 - m_4) + O_p(n^{-3/2}).
\]

The signed root of the conditional profile likelihood ratio statistic for \( \theta = \mu_1 \) has the expansion

\[
\bar{R} = V + \frac{1}{2} (m_1 - m_4) + \frac{1}{3} m_1 V^2 + O_p(n^{-1}),
\]

and the signed root of the likelihood ratio statistic is

\[
R = V + \frac{1}{3} m_1 V^2 + O_p(n^{-1}).
\]
Therefore, \( \bar{R} = R + \frac{1}{2}(m_1 - m_4) + O_p(n^{-1}) \), \( E(\bar{R}) = \frac{1}{2}m_1 - \frac{1}{2}m_4 + O_p(n^{-1}) \), and \( R - (\frac{1}{2} \hat{m}_1 - \frac{1}{2} \hat{m}_4) = \bar{R} + z_0 + O_p(n^{-1}) \). The confidence limits obtained from the standard normal approximation to the distribution of \( \bar{R} - (\frac{1}{2} \hat{m}_1 - \frac{1}{2} \hat{m}_4) \) are then second-order accurate. Thus, the ABC confidence limits agree with limits derived from the conditional profile likelihood to second order.

For models indexed by a scalar parameter, Welch and Peers (1963) have discussed the use of Bayesian procedures to construct approximate confidence limits. In particular, they showed that if the prior distribution for a parameter of interest \( \theta \) is chosen proportional to the square root of the expected Fisher information for \( \theta \), then the Bayesian posterior \( 1 - \alpha \) quantile for \( \theta \) is an approximate upper \( 1 - \alpha \) confidence limit having coverage probability of \( 1 - \alpha + O(n^{-1}) \). Thus, the \( 1 - \alpha \) posterior quantile is second-order accurate. We now show that it is second-order correct as well.

For the exponential family model (4.2), when \( p = 1 \) and the parameter of interest is \( \theta = t(\mu) \), we have
\[
\sigma^2 = \psi^{(2)}(t^{(1)})^2, \quad m_1 = \psi^{(3)}(t^{(1)})^3 / \sigma^3, \quad m_2 = m_3 = \psi^{(2)}(t^{(2)}) / \sigma,
\]
\[
z_0 = a = \frac{1}{6}m_1 = \frac{1}{6} \text{sgn}(t^{(1)}) \psi^{(3)}(\psi^{(2)})^{-3/2}, \quad \text{curv} = 0,
\]
where \( t^{(k)} = d^k t(\mu) / d\mu \) and \( \psi^{(2)} = d^2 \psi(\eta) / d\eta^2 \). Now the prior density prescribed by Welch and Peers (1963) is proportional to \( |t^{(1)}|^{-1/2} \). Their formula (17) provides a general expansion for the confidence limit obtained using their approach. For the present problem, this expansion becomes
\[
\hat{\theta} + z_0 \hat{\sigma} + \frac{1}{6} (z_0^2 + 2)(3 \psi^{(3)} t^{(2)} + 2 \psi^{(3)} t^{(1)} / \psi^{(2)}) - \frac{1}{2} (2 \psi^{(2)} t^{(2)} + \psi^{(3)} t^{(1)} / \psi^{(2)}) + O_p(n^{-3/2})
\]
\[
= \hat{\theta} + \hat{\sigma} \{ z_0 + \frac{1}{6} m_1 + \frac{1}{3} m_1 + \frac{1}{2} m_3 z_0^2 \} + O_p(n^{-3/2}),
\]
and hence the Welch and Peers approximate confidence limit is second-order correct.

In summary, we have shown that second-order accurate confidence limits obtained using the profile likelihood, the Barndorff-Nielsen and Cox approximate conditional likelihood, the Cox and Reid conditional profile likelihood, and the Welch and Peers Bayesian approach are all second-order correct according to the present definition based on the distribution of the quantity \( V = (\hat{\theta} - \theta) / \hat{\sigma} \). The ABC limits introduced in this paper therefore agree with all of these limits to second order.

REFERENCES


DiCiccio, T. J., and Martin, M. A (1990), "Approximations of Marginal Tail Probabilities for a Class of Smooth Functions with Applications to Bayesian and Conditional Inference", Technical Report, Stanford University, Department of Statistics.


APPENDIX: S Programs

Here is the program "abc" for calculating ABC interval endpoints. It is written in the language S, Becker, Chambers, and Wilks (1988). The letters at left refer to the annotated comments following the program listing. The arguments tt, y, S, etahat, mu are explained in the second paragraph of Section 3. Section 3 gives the exact forms of y, S, etahat, and mu applying to the most familiar exponential families.

```s
"abc" <- function(tt, y, S, etahat, mu, lambda = 0.01, alpha = c(0.025, 0.05, 0.1, 0.16)) {
  p <- len(y)
  I <- diag(p)
  
  # calculate thetahat, ehat, dhat, and sighat
  thetahat <- tt(y)
  ehat <- numeric()
  for (j in 1:p) {
    lam <- lambda * S[j, j]^0.5
    delta <- I[, j]
    ehat[j] <- (tt(y + lam * delta) - tt(y - lam * delta))/(2 * lam)
  }
  dhat <- as.vector(S * ehat)
  sighat <- sqrt(ehat * S * ehat)

  # calculate acceleration a
  lam <- lambda/sighat
  a0 <- sum(ehat * mu(etahat))
  a1 <- sum(ehat * mu(etahat + lam * ehat))
  a2 <- sum(ehat * mu(etahat - lam * ehat))
  a <- (a1 - 2 * a0 + a2)/(lam^2 * 6 * sighat^3)

  # calculate curvature curv
  curves <- numeric()
  eig <- eigen(S)
  evals <- (eig$values)^0.5
  evvecs <- eig$veectors
  for (j in 1:p) {
    c1 <- tt(y + lambda * evals[j] * evvecs[, j])
    c2 <- tt(y - lambda * evals[j] * evvecs[, j])
    curves[j] <- (c1 - 2 * thetahat + c2)/lambda^2
  }
  delta <- dhat/sighat
  curv0 <- (tt(y + lambda * delta) - 2 * thetahat + tt(y - lambda * delta))/lambda^2
  curv <- (sum(curves) - curv0)/(2 * sighat)

  # calculate zla
  z0 <- qnorm(2 * pnorm(a) * pnorm(-curv))
  a1 <- c(alpha, rev(1 - alpha))
  za <- qnorm(a1)
  z0a <- (z0 + za)/(1 - a * (z0 + za))
  zla <- z0a + a * z0a^2

  # calculate endpoints
  standard <- thetahat + sighat + za
  ABC <- numeric()
  dhat <- dhat/sighat
  for(j in 1:len(zla)) {
    ABC[j] <- tt(y + dhat[j] * zla[j])
  }
  ABCquad <- thetahat + sighat + zla + 0.5 * curv0 * zla^2

  llims <- cbind(1, ABC, ABCquad, standard)
  limnames(llims) <- list(NULL, c("alpha", "ABC", "ABCquad", "Standard"))

  # output in list form
  ls <- list()
  ls$llims <- llims
  ls$thetahat.sighat <- c(thetahat, sighat)
  ls$a.z0 <- c(a, z0)
  ls$curv.nonlin <- c(curv, curv0/(2 * sighat))
  ls
```

28
Comments:

A. The value of lambda controls the increment used in evaluation of numerical derivatives; lambda= .01 was used in our examples. The ABC endpoints are evaluated for each value of alpha and 1−alpha.

B. J is a p × p identity matrix.

C. thetab = \hat{\theta}, sigmab = \hat{\sigma}, ehat and dhat are the least favorable directions $\hat{\epsilon}$ and $\hat{\delta}$ shown in Figure 2.

C'. The actual increment "lam" used in evaluating ehat,j is lambda times $sd(\gamma_j)$; similarly for the other numerical derivatives.

D. Formula (2.30).

E. Formula (2.31).

F. Eigenvalue and eigenvectors of $S$; gives vectors $C_j$ in (2.31).

G. Formula (2.32); pnorm, qnorm are functions $\Phi, \Phi^{-1}$; can also use (2.33).

H. Formula (2.23).

I. Formula (1.1).

J. Formula (2.21).

K. Quadratic formula (2.26).

L. "nonlin" is the ratio of quadratic to linear coefficients in (2.26), $\delta' \hat{\delta} / 2\hat{\sigma}^3$; it measures the non-linearity of the functions $t(\mu)$.

The program Sdiff is useful for situations where the covariance matrix $S = \Sigma(\mu)$ is not easily available. Sdiff calculates $S$ by the numerical differentiation of $mu(\eta)$ at $\eta = \hat{eta}$, (3.21):

```r
"Sdiff" <-
function(eta, mu, lambda=.001, SO) {
  # differentiate mu(eta) to get covariance S
  p <- length(eta)
  I <- diag(p)
  S <- matrix(0, p, p)
  for(j in 1:p) {
    if(is.null(SO))
      lam <- lambda * (abs(eta[j]) + 1e-10)
    else lam <- lambda / SO[j, j]^0.5
    S[,j] <- (mu(eta+lam*I[,j]) - mu(eta-lam*I[,j]))/(2*lam)
  }
}
```

A. The arguments etahat, mu, and lambda are the same as for the program abc; SO is an optional preliminary guess for $S$.

B. If SO is not supplied originally, it is a good idea to run Sdiff a second time, with SO equal to the first estimate of $S$.

In the logistic regression example of Figure 5, we need to calculate $tt(y + del) = s(\hat{\eta}(y + del))$ for various choices of the incremental vector del. The difficult part of the calculation is $\hat{\eta}(y + del)$, since here the function $\eta(\mu)$ does not have a closed form. The program etalogit calculates $\hat{\eta}(y + del)$ using Newton-Raphson updating, (3.14).
It is assumed that the following quantities are available in memory; \( n \), the \( N \times 1 \) vector of sample sizes in (3.9); \( X = (x_1, x_2, \ldots, x_N) \), the \( p \times N \) matrix of covariate vectors (3.10); \( y = Xz \), the \( p \times 1 \) sufficient vector; \( Si = S^{-1} \), the \( p \times p \) inverse of the covariance matrix \( S = \sum_{i=1}^{n} x_i n_i \hat{\pi}_i (1 - \hat{\pi}_i) x_i' \); and finally \( J \), one less than the number of iterations of (3.14) performed. (\( J = 3 \) in the example of Figure 5.)