The Canadian Journal of Statistics
Vol. 37, No. 1, 2009, Pages 1–21
La revue canadienne de statistique

Three enigmatic examples and inference from likelihood

D. A. S. FRASER[1], A. WONG[2]* and Y. SUN[1]

[1] *Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3*
[2] *Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada M3J 1P3*

Key words and phrases: Asymptotics; Behrens–Fisher; Box and Cox; conditioning; Cox measuring instrument; large sample; likelihood; simulations.

MSC 2000: Primary 62F03; secondary 62G32.

Abstract: Statistics has many inference procedures for examining a model with data to obtain information concerning the value of a parameter of interest. If these give different results for the same model and data, one can reasonably want a satisfactory explanation. Over the last eighty years, three very simple examples have appeared intermittently in the literature, often with contradictory or misleading results; these enigmatic examples come from Cox, Behrens, and Box & Cox. The procedures in some generality begin with an observed likelihood function, which is known to provide just first order accuracy unless there is additional information that calibrates the parameter. In particular, default Bayes analysis seeks such calibration in the form of a model-based prior; such a prior with second order accuracy is examined for the Behrens problem, but none seems available for the Box and Cox problem. Alternatively, higher-order likelihood theory obtains such information by examining likelihood at and near the data and achieves third order accuracy. We examine both Bayesian and frequentist procedures in the context of the three enigmatic examples; simulations support the indicated accuracies. *The Canadian Journal of Statistics* 37: 1–21; 2009 © 2009 Statistical Society of Canada

Résumé: La Statistique offre plusieurs procédures permettant d'inférer la valeur d'un paramètre d'intérêt à l'aide d'un modèle et de données. Si des procédures utilisant le même modèle et les mêmes données conduisent à des résultats différents, il est légitime d'exiger une explication satisfaisante. Au cours des quatre-vingt dernières années, trois exemples apparaissant sporadiquement dans la littérature ont donné des résultats souvent trompeurs ou contradictoires. Ces exemples énigmatiques par Cox (1958), Behrens (1929) et Box & Cox (1964) sont généralement basés sur une fonction de vraisemblance observée dont la précision est du premier ordre, sauf en présence d'information additionnelle permettant de calibrer le paramètre. En analyse Bayésienne, par exemple, cette calibration se présente sous la forme d'une loi à priori. Une telle loi dont la précision est du deuxième ordre (Ghosh & Kim; 2001) est examinée pour le problème de Behrens, mais aucun équivalent ne semble possible pour le problème de Box & Cox. De son côté, la théorie de la vraisemblance d'ordre supérieur déduit cette information en examinant la fonction de vraisemblance aux données et près de celles-ci, permettant une précision du troisième ordre. Les auteurs étudient des procédures Bayésiennes et fréquentistes dans le contexte de ces trois exemples énigmatiques; des simulations confirment les précisions annoncées. *La revue canadienne de statistique* 37: 1–21; 2009 © 2009 la Société statistique du Canada

1. INTRODUCTION

Three statistical problems, quite elementary in appearance, have been examined many times in the literature, from various pragmatic and methodological approaches. The examples are the Cox measuring instrument in Welch (1939) and in Cox (1958); the Behrens–Fisher two sample

* *Author to whom correspondence may be addressed.*
E-mail: august@mathstat.yorku.ca

problem in Behrens (1929) and in Ghosh & Kim (2001); and the Box and Cox problem in Box & Cox (1964) and in Chen, Lockhart & Stephens (2002); the examples are simple to describe and typically involve just normally distributed variables.

The pragmatic approach would be illustrated by taking just the mean and standard deviation of a convenient statistic such as a Student departure for the Behrens–Fisher problem, to the fine tuning of an approximate degrees of freedom for that statistic as with Welch's (1947) insightful proposal.

Methodological approaches are illustrated by the use of likelihood in the original Bayes manner or the use of a likelihood ratio calculation as with the Behrens–Fisher and Box and Cox problems. The results are sometimes in close agreement and sometimes quite different, We thus refer to the examples as enigmatic. Perhaps the enigma lie deeper. Do we have just a tool bag of methods to explore with? Or is there some unifying way of extracting all or almost all the available statistical information from a model data combination?

For the simple normal model with unknown location and scaling, almost all resolutions would lead to the familiar Student quantity with the usual Student distribution. How then can two normal models as with the Cox example and the Behrens–Fisher example produce such difficulties for resolution? Or how can the presence of reexpressed normal variables as with the Box and Cox example produce such difficulties?

Daniels (1954) working from applied mathematics showed that probability calculations can be much more accurate through the use of a cumulant generating function. This did have some substantial messages for statistics but these did not emerge for some 25 years until Barndorff-Nielsen & Cox (1979) presented various statistical applications for the cumulant generating function approach. A substantial limitation however was the need for the cumulant generating function. Of course, cumulant generating functions are typically implicit in full exponential family models and such models were indeed highlighted by Barndorff-Nielsen & Cox (1979), but with notation in the more familiar likelihood form.

The restriction to exponential family models with available sufficiency meant that even location-scale models with nonnormal error are not covered. Subsequently Barndorff-Nielsen (1986) examined log statistical models and with third order probability calculations extended the cumulant generating function approach. Again there were limitations; the analysis typically required that the variable and the parameter have the same effective dimension, if not directly, then by ancillary conditioning.

More recently, a general extension builds on approximate ancillarity derived from model quantile continuity. We address the quite general availability of this higher-order likelihood approach, a natural follow through from the cumulant generating function approach.

In the next section, we briefly describe the examples and survey the likelihood based methods that are applicable. We will see that the likelihood function contains the accuracy of a cumulant generating function for the Daniels' type calculation of probabilities, and that parameter information just needs to be extracted by approximate methods that are highly accurate.

2. THREE SIMPLE EXAMPLES AND LIKELIHOOD INFORMATION

In this section we briefly describe the three simple statistical problems, which have little structure beyond that of a sample from a normal distribution. We also briefly survey the higher-order methods for extracting the essential information from a likelihood function. Then in later sections we survey the application of the methods to the examples.

As the first example, consider the measuring instrument context (Cox, 1958), which has close parallels in an example used in Welch (1939). Cox (1958) considered two measuring instruments, both unbiased and normal but with different standard deviations, say $\sigma_1 = 100\sigma_0$ and $\sigma_2 = \sigma_0$; the context also included an equally likely random choice of which instrument to use to make a single measurement of a scalar parameter θ. This random choice may seem artificial but it has clear parallels embedded in almost all modelling contexts. Cox pondered the appropriate sample space for inference and then developed such in terms of conditioning on an ancillary statistic (Fisher, 1925, 1934, 1935a,b). By contrast, Welch had focused on maximizing power for given test size. We discuss this example in Section 3 and make direct use of continuity as to how the parameter moves the distribution of individual coordinates; we do not however address the power approach.

As the second example, consider the Behrens (1929)–Fisher (1935a,b) problem. This involves a sample of n_1 values from a Normal (μ_1, σ_1^2) and a sample of n_2 values from a Normal (μ_2, σ_2^2). The two sampling models are independent, and the interest parameter is the difference in means $\delta = \mu_1 - \mu_2$. This does involve just two normal samples yet no generally acceptable or definitive procedure has evolved in the literature. Fisher (1935a,b) following Behrens (1929) obtained a confidence distribution for μ_1 and a confidence distribution for μ_2, and then convolved these to obtain a distribution for δ. At that time, the confidence distributions were called fiducial distributions, and they were based on the now familiar confidence pivotal inversion. Both frequentists and Bayesians took exception to the procedure, the former to the combining of two confidence distributions and the latter to the confidence inversion itself. Jeffreys (1961) proposed the prior $\sigma_1^{-1}\sigma_2^{-1}$ relative to $d\mu_1\, d\mu_2\, d\sigma_1\, d\sigma_2$, which is the product of the familiar right invariant measures for the separate location-scale groups; this differs from the usual Jeffreys prior $\sigma_1^{-2}\sigma_2^{-2}$ which corresponds to the left invariant measures. Ghosh & Kim (2001) used a second order asymptotic argument and frequentist properties to develop the prior

$$\sigma_1^{-3}\sigma_2^{-3}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

For either approach, a joint posterior distribution for μ_1, μ_2, σ_1^2, and σ_2^2 can be obtained, and then the marginal posterior distribution for $\mu_1 - \mu_2$. Ghosh & Kim compared the results from the new prior and from the right invariant Jeffreys prior by examining the frequency of correct statements using related posterior intervals. For larger sample sizes and most smaller sample sizes the new second order prior gave a substantially higher frequency of correct statements. The example is discussed in Section 4 using Bayesian and likelihood approaches.

As the third example, consider the Box & Cox (1964) problem: a response variable suitably expressed has a linear model $X\beta + \sigma z$ where X is $n \times r$ matrix of full column rank r, β is an $r \times 1$ vector of regression coefficients, z is an $n \times 1$ sample from a standard normal distribution, and σ is a positive scalar. A familiar method of response re-expression is provided by the power transformations: for a positive coordinate y_i, the re-expressions y_i^λ with λ not equal to zero form a transformation group; the limit case $\lambda = 0$ can be linked to the transformation $\log(y)$ provided X includes the 1-vector. For notation, let $y_1^\lambda, \ldots, y_n^\lambda$ designate the vector of transformed responses; the model then can be written as

$$y = (X\beta + \sigma z)^{1/\lambda}.$$

Box & Cox (1964) obtained maximum likelihood and maximum Bayesian posterior estimates for λ, and then used the estimated value for λ as a fixed value for a regression analysis of the remaining parameters; this latter step is often called a "plug-in analysis." Various objections were raised by Bickel & Doksum (1981) to the direct use of the estimated value; and these were answered by Box & Cox (1982). Chen, Lockhart & Stephens (2002) discussed the choice of the parameter to be estimated while focussing on stability of the corresponding estimation analysis and gave preference to a ratio β/σ of a regression parameter β to error standard deviation σ; they then provided simulations to demonstrate the stability. In the discussion, McCullagh raised questions whether the choice β/σ was "physically meaningful" and mentioned the notion of a natural subparameter (McCullagh, 2002); the resulting claim was that β/σ did not represent a characteristic of the context being investigated and was rather an artefact of the model and thus inappropriate to pursue. The example is discussed in Section 6 using present higher-order likelihood analysis methods.

The methodological approaches to the examples cover a broad spectrum and typically make direct use of the observed likelihood function $L^o(\theta) = cf(y^o;\theta)$ where $f(y;\theta)$ is the model in density form, y^o is the observed data point, and c is an arbitrary positive constant. Exceptions are the pragmatic choice of the degrees of freedom (Welch, 1947) for the Behrens–Fisher problem and the Welch (1939) use of optimality and fixed test size for the Cox (1958) problem; we do not address these here.

Bayes (1763) introduced the direct use of the full observed likelihood function $L^o(\theta)$ as the primary ingredient for statistical inference; this coupled the observed likelihood $L^o(\theta)$ with a weight function $\pi(\theta)$ to give an evaluation $c\pi(\theta)L^o(\theta)$ on the parameter space. In the context he examined, the model was specialized and had location form which would now be written $f(y-\theta)$ and the weight function was constant reflecting the translation invariance of the model; this use of likelihood was long before its formal recognition (Fisher, 1922) and the supporting argument invoked conditional probability even though $\pi(\theta)$ in context was not describing probability; the approach can be viewed as innovative, exploratory and much ahead of its time. If we have interest in a scalar component parameter $\psi(\theta)$, we might then reasonably calculate the posterior survival value

$$\int_{\psi(\theta)=\psi}^{\infty} c\pi(\alpha)L^o(\alpha)\,\mathrm{d}\alpha$$

with integration over the parameter region having $\psi(\theta)$ values larger than some interest value ψ; for this various first order and higher-order approximations are surveyed in Bédard et al. (2008).

Various characteristics of the observed likelihood function can focus on scalar characteristics such as the maximum likelihood value $\hat{\theta}^o = \hat{\theta}(y^o)$ or the score $\ell_\theta(\theta; y^o) = s(\theta)$ which is the slope of the log-likelihood function at a tested value θ; scaling of these can be obtained from several sources but the curvature or the Hessian at the maximum likelihood value has strong support and is called the observed information

$$j_{\theta\theta'}(\hat{\theta}^o) = -\ell_{\theta\theta}(\hat{\theta}^o; y^o) = -\ell_{\theta\theta}(\hat{\theta}^o) = -\left.\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}\ell(\theta)\right|_{\hat{\theta}^o}.$$

Thus for the maximum likelihood departure from an interest value θ in the scalar case, we have

$$q(\theta) = j_{\theta\theta'}(\hat{\theta}^o)^{1/2}(\hat{\theta}^o - \theta)$$

which is first order standard normal from an asymptotic view.

The maximum likelihood departure is directly affected by a change in the parameterization. A familiar departure measure that has invariance to reparameterization is the signed log-likelihood ratio, sometimes referred to as the directed deviance (Barndorff-Nielsen & Cox, 1994). For a scalar interest parameter $\psi = \psi(\theta)$ it has the form

$$r(\psi) = r(\psi, y^o) = \mathrm{sgn}(\hat{\psi}^o - \psi)\left[2\left\{\ell(\hat{\theta}^o) - \ell(\hat{\theta}^o_\psi)\right\}\right]^{1/2}, \qquad (1)$$

where $\hat{\theta}^o_\psi$ is the maximum likelihood value when the parameter is restricted to values satisfying $\psi(\theta) = \psi$; from the asymptotic view $r(\psi)$ is also first order standard normal.

An improvement on the preceding q which covers the more general case of a scalar component $\psi(\theta)$ takes the form

$$q = (\hat{\psi}^o - \psi)\left\{\frac{|j_{\theta\theta'}(\hat{\theta}^o)|}{|j_{\lambda\lambda'}(\hat{\theta}^o_\psi)|}\right\}^{1/2} \tag{2}$$

where $j_{\theta\theta'}(\hat{\theta}^o) = -(\partial^2/\partial\theta\partial\theta')\ell(\theta)|_{\theta=\hat{\theta}^o}$, and $j_{\lambda\lambda'}(\hat{\theta}^o_\psi) = -(\partial^2/\partial\lambda\partial\lambda')\ell(\theta)|_{\theta=\hat{\theta}^o_\psi}$, $\hat{\psi}^o = \psi(\hat{\theta}^o)$ and λ is a complementing nuisance parameter having $\theta = (\psi, \lambda')'$. When the j matrices are too complicated in form, they can be evaluated numerically by fine differencing provided the likelihood function is calculated with high accuracy. Again this revised q is first order standard normal but its statistical quality can be seriously degraded in the presence of nuisance parameters. Note that (1) and (2) give approximate p-values $\Phi(r)$ and $\Phi(q)$, where $\Phi(\cdot)$ is the standard normal distribution function, but the first based on the signed log-likelihood ratio is widely viewed as more reliable.

The full third order use of the cumulant generating type information contained in the likelihood function and mentioned in Section 1 extends from cumulant generating functions themselves, to the exponential model context, to approximate exponential models, to exact conditioning, and then to approximate conditioning. Key formula for this make use of a combination of r and q type quantities inherited from contexts with cumulant generating functions; the formula has two versions, one from Lugannani & Rice (1980) and the other from Barndorff-Nielsen (1986), the latter being

$$p(\psi) = \Phi(r^*) = \Phi\left\{r - r^{-1}\log\left(\frac{r}{q}\right)\right\} \tag{3}$$

where r^* is implicitly defined, r is the signed log-likelihood ratio (1) and q given in (4) is a modification of (2) that uses an intrinsic reparameterization $\varphi(\theta)$ which establishes the analytic link to the cumulant generating function.

The needed maximum likelihood based departure q works within the reparameterization $\varphi(\theta)$:

$$q = q(\psi) = q(\psi; y^o) = \mathrm{sgn}(\hat{\psi}^{\mathrm{o}} - \psi)|\chi(\hat{\theta}^{\mathrm{o}}) - \chi(\hat{\theta}^{\mathrm{o}}_\psi)|\left\{\frac{|\mathrm{j}_{\varphi\varphi'}(\hat{\theta}^{\mathrm{o}})|}{|\mathrm{j}_{(\lambda\lambda')}(\hat{\theta}^{\mathrm{o}}_\psi)|}\right\}^{1/2}. \tag{4}$$

For this $\chi(\theta)$ is a special linear combination of the coordinates of the reparameterization $\varphi(\theta)$ and the two information matrices $j_{\varphi\varphi'}(\hat{\theta}^o)$ and $j_{(\lambda\lambda')}(\hat{\theta}^o_\psi)$ are just observed informations or Hessians for the full and for the nuisance parameter, but calculated in the φ parameterization. Some details on the rescaling to the φ parameterization are given in Appendix together with some technical and expository references. The computations involve just first and second derivatives of likelihood and are readily available by calculus through Mathematica or Maple, or by numerical differencing on a fine scale using high precision calculations (Davision, Fraser & Reid, 2006).

The reparameterization $\varphi(\theta)$ is to provide a stand-in for the canonical parameter of an exponential model and thus to link it to the usual Daniels type arguments using cumulant generating functions. The reparameterization is available as the observed gradient of the log-likelihood function,

$$\varphi'(\theta) = \ell_V(\theta; y)|_{y^o} = \left(\frac{\mathrm{d}}{\mathrm{d}v_1}, \cdots, \frac{\mathrm{d}}{\mathrm{d}v_p}\right)\ell(\theta; y)\bigg|_{y^o}, \tag{5}$$

calculated in directions $V = (v_1, \ldots, v_p)$ that describe the approximate conditioning that extends the methodology from the sufficient statistic context to the general asymptotic context; the subscript V is to indicate that directional derivatives are taken in the directions recorded in the matrix V, for example, $(\mathrm{d}/\mathrm{d}v)\ell(\theta; y) = (\mathrm{d}/\mathrm{d}t)\ell(\theta; y + tv)|_{t=0}$.

The directions $(v_1, \ldots, v_p)$ forming the matrix V are tangents to an approximate ancillary and are obtained as a gradient of the coordinate quantile functions. Let $y_i = y_i(u_i, \theta)$ be the quantile function corresponding to the coordinate distribution function $u_i = F_i(y_i, \theta)$; for example with the Normal (μ, σ^2), the distribution function is $\Phi\{(y - \mu)/\sigma\}$ and the quantile is then the inverse function $y = \mu + \sigma z$ using $u = \Phi(z)$. The direction $V = (v_\alpha) = (v_{i\alpha})$ are given by $v_{i\alpha} = \partial y_i/\partial\theta_\alpha|_{(\hat{u}_i^o, \hat{\theta}^o)}$ where $\hat{u}_i^o = F_i(y_i^o, \hat{\theta}^o)$; see Fraser & Reid (1995, 2001).

The p-value for a parameter ψ as given by $p(\psi)$ in (3) is pivotal and from the sampling viewpoint it has the Uniform(0,1) distribution with third order accuracy. It can analytically be viewed as a definitive p-value to third order. And it leads to any choice of confidence interval or confidence bound. For example the 95% confidence lower bound is

$$\hat{\psi}_{0.95} = p^{-1}(0.95)$$

and the central 95% confidence interval is

$$(\hat{\psi}_{0.975}, \hat{\psi}_{0.025}) = (p^{-1}(0.975), p^{-1}(0.025));$$

this is just standard confidence or pivotal inversion (Fisher, 1930, 1935a,b; Neyman, 1937). The function $p(\psi)$ as given in (3) is called here the p-value function for ψ.

3. COX MEASURING INSTRUMENT EXAMPLE

The Cox (1958) measuring instrument example discussed briefly in Section 1 can be expressed in a simple form: y is distributed as Normal (θ, σ_a^2), and a is Bernoulli (1/2). The observable variable (y, a) has sample space $R \times \{1, 2\}$.

Consider now an observed data point (y^o, a^o). The observed likelihood function is

$$L^o(\theta) = \exp\left\{-\frac{(y^o - \theta)^2}{2\sigma_{a^o}^2}\right\}$$

which is location normal in shape with scaling given by the data-identified standard deviation σ_{a^o}. The corresponding observed log-likelihood function is

$$\ell(\theta) = -\frac{(y^o - \theta)^2}{2\sigma_{a^o}^2};$$

this clearly reflects the use of the observed precision as indicated by the scaling σ_a^2 coming from the actual instrument that made the measurement.

The model has the translation invariance central to Bayes (1763) and the corresponding default prior is $\pi(\theta) = c$. This gives the posterior that θ is Normal $(y^o, \sigma_{a^o}^2)$.

The decision-theoretic approach supported by Welch (1939) would ignore that the scaling had been observed and was known; this is seemingly contradictory to clear objectives of statistical inference. The result is that the two possible values for a lead correspondingly to separate critical values for y. This is a trade-off between the two p-values to achieve some overall optimality.

From the higher-order likelihood approach we would calculate the gradient of the log-likelihood function at the observed data and obtain

$$\ell_y(\theta) = \varphi(\theta) = \frac{\theta - y^o}{\sigma_{a^o}^2}$$

which is affine in θ; this calculation of gradient describes what small parameter change at the data point does to log-likelihood. An affine transformation on $\varphi(\theta)$ has no effect on (1) or (4); thus it suffices to take φ to be just the given θ as expected. Then routinely from (1) and (5) we obtain

$$r = \frac{y^o - \theta}{\sigma_{a^o}}, \qquad q = \frac{y^o - \theta}{\sigma_{a^o}};$$

and then from (3) we have

$$\begin{aligned} p(\psi) &= \Phi\left\{\frac{y^o - \theta}{100\sigma_0}\right\} \quad \text{if } a = 1, \\ &= \Phi\left\{\frac{y^o - \theta}{\sigma_0}\right\} \quad \text{if } a = 2. \end{aligned}$$

This is just the p-value based on the model for the measuring instrument that actually made the measurement, a very natural result. This of course agrees with Cox (1958) who conditions on the ancillary $a = a^o$ as suggested in Fisher (1925, 1934, 1935a,b). For some recent discussion of conditioning see Fraser (2004). The Bayesian and frequentist approaches agree and run counter to the optimality approach supported by Welch (1939).

4. BEHRENS–FISHER EXAMPLE

If a sample from a normal distribution can be viewed as providing a simple primal statistical problem, then surely samples from two separate normal distributions should come a close second. But the almost eight decades since the original Behrens (1929) paper indeed suggest otherwise. Let $(y_{11}, \ldots, y_{1n_1})$ be an independent sample from the Normal (μ_1, σ_1^2) and $(y_{21}, \ldots, y_{2n_2})$ be an independent sample from the Normal (μ_2, σ_2^2); and suppose we are interested in $\delta = \mu_1 - \mu_2$. Let $\bar{y}_1, \bar{y}_2$ be the sample means and s_1^2, s_2^2 be the sample variances.

From a pragmatic frequentist viewpoint, a fairly natural quantity for assessing δ is readily available as

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}}, \tag{6}$$

which is the departure of the estimate $\bar{y}_1 - \bar{y}_2$ from the interest parameter δ, standardized by its estimated standard deviation. At issue however is the full statistical calibration of this quantity: it does not have a distribution free of the nuisance parameters. Both Behrens (1929) and Fisher (1935a,b) produced a distribution for assessing δ but the derivations were not generally accepted. Welch (1947) noted the close connection to the ordinary t-statistic and chose a degrees of freedom that closely approximated the denominator to a scaled Chi variable, and then recommended the corresponding Student distribution.

From the default Bayesian viewpoint mentioned in Section 1, we have the left and right Jeffreys priors and the recent second order Ghosh & Kim (2001) prior.

Now consider the higher-order likelihood approach as described in Section 2. The observed log-likelihood function is

$$\ell(\theta) = -\frac{1}{2\sigma_1^2}\{n_1(\bar{y}_1 - \mu_1)^2 + S_1^2\} - \frac{1}{2\sigma_2^2}\{n_2(\bar{y}_2 - \mu_2)^2 + S_2^2\} - \frac{n_1}{2}\log\sigma_1^2 - \frac{n_2}{2}\log\sigma_2^2,$$

where S_1^2 and S_2^2 are the sums of squares of residuals in the first and second samples. The effective sample space involves $(\bar{y}_1, \bar{y}_2, S_1^2, S_2^2)$ and has dimension 4, equal to that of the parameter. The model itself is exponential and one version of the canonical parameter is

$$\varphi(\theta) = \varphi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \varphi = \left(\frac{\mu_1}{\sigma_1^2}, \frac{1}{\sigma_1^2}, \frac{\mu_2}{\sigma_2^2}, \frac{1}{\sigma_2^2}\right),$$

which we take as a row vector. This is a case where the cumulant generating function is implicitly available and the higher-order likelihood approach just mechanizes the calculations of the p-value.

If we follow the higher-order likelihood procedure and calculate the sample space gradient of the observed log-likelihood function at the observed data we would obtain just an affine function of the row vector above and it would be an equivalent canonical parameterization. This happens quite generally: if a model is a full exponential model, the sample space gradient procedure just extracts an exponential canonical parameterization; and an affine change in the parameterization does not affect the calculated p-value; in effect it works from the implicit cumulant generating function.

The maximum likelihood values are almost immediate. For the full parameter θ the maximum likelihood value $\hat{\theta}$ is just the combination of the individual sample maximum likelihood values: $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2) = (\bar{y}_1, \bar{y}_2, S_1^2/n_1, S_2^2/n_2)$. For the constrained maximum likelihood value, the exact functional form seems not available, but it is easily obtained numerically.

5. BEHRENS–FISHER SIMULATIONS

We use simulations to evaluate the inference procedures for the Behrens–Fisher problem targeted on the difference $\delta = \mu_1 - \mu_2$ in the population means. The inference procedures discussed in Section 4 include the posterior distribution based on right invariant Jeffreys prior, the posterior distribution based on the second order Ghosh & Kim (2001) prior, the confidence distribution based on the Welch test approximation, the confidence distribution based on the signed log-likelihood ratio r, and the confidence distribution based on the third order adjustment r^*. And for each procedure, we focus on central $(1 - \alpha)100\%$ intervals $(\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2})$ where $\hat{\theta}_\gamma$ is the γ quantile of a proposed inference distribution. And to enable direct comparison with Ghosh & Kim (2001) results, we take $\alpha = 10\%$.

While there are four parameters in the full model, the evaluation of the procedures only needs the ratio σ_1^2/σ_2^2 of the variances and of course the sample sizes n_1, and n_2. In particular, without loss of generality, we choose $\mu_1 = 2$, $\mu_2 = 0$, and $\sigma_2^2 = 1$, and then follow Ghosh & Kim with $\sigma_1^2 = 2$, and $\sigma_1^2 = 4$ and with various choices of n_1 and n_2, as indicated in Tables 1a and 1b.

We first generate $N = 10{,}000$ instances of a sample of n_1 for y_1 and of a sample of n_2 for y_2, for a given choice of the parameters as just described. In each instance, we calculate the central 90% parameter quantile interval. Table 1 reports the proportion of cases where the true value falls below the lower limit of the interval, and the proportion of cases below the upper limit of the interval; these should have the target values 0.05 and 0.95; we also record the simulation standard deviations which is calculated as $\sqrt{p(1-p)/N}$, where p is the target value.

TABLE 1a: For $\mu_1 = 2.0$, $\mu_2 = 0.0$, $\sigma_2^2 = 1.0$ without loss of generality and for various $n_1 \geq n_2$, σ_1^2, we record the proportion of the 10,000 cases where the true δ is less than the lower limit and less than the upper limit of the 90% central interval.

			$\sigma_1^2 = 2.0$		$\sigma_1^2 = 4.0$	
n_1	n_2	Method	<Lower limit	<Upper limit	<Lower limit	<Upper limit
		Target	0.0500	0.9500	0.0500	0.9500
		Sim SD	0.0022	0.0022	0.0022	0.0022
2	2	Jeffreys	0.0094	0.9924	0.0107	0.9908
		Ghosh & Kim	0.0199	0.9841	0.0221	0.9812
		slr	0.1322	0.8718	0.1396	0.8636
		Welch	0.0293	0.9701	0.0349	0.9664
		Third order	0.0274	0.9731	0.0304	0.9677
20	2	Jeffreys	0.0385	0.9681	0.0295	0.9745
		Ghosh & Kim	0.1097	0.8970	0.0924	0.9096
		slr	0.1286	0.8685	0.1117	0.8868
		Welch	0.0829	0.9191	0.0724	0.9281
		Third order	0.0683	0.9300	0.0646	0.9332
3	2	Jeffreys	0.0133	0.9895	0.0128	0.9899
		Ghosh & Kim	0.0279	0.9745	0.0287	0.9733
		slr	0.1072	0.8954	0.1024	0.8989
		Welch	0.0356	0.9662	0.0364	0.9657
		Third order	0.0313	0.9664	0.0331	0.9655
7	5	Jeffreys	0.0348	0.9651	0.0361	0.9631
		Ghosh & Kim	0.0474	0.9525	0.0458	0.9525
		slr	0.0716	0.9285	0.0704	0.9310
		Welch	0.0494	0.9504	0.0479	0.9490
		Third order	0.0517	0.9488	0.0528	0.9505
15	10	Jeffreys	0.0410	0.9509	0.0411	0.9487
		Ghosh & Kim	0.0466	0.9435	0.0467	0.9447
		slr	0.0663	0.9434	0.0652	0.9435
		Welch	0.0468	0.9432	0.0472	0.9444
		Third order	0.0572	0.9528	0.0561	0.9524
20	15	Jeffreys	0.0486	0.9608	0.0486	0.9588
		Ghosh & Kim	0.0531	0.9549	0.0520	0.9550
		slr	0.0525	0.9402	0.0509	0.9420
		Welch	0.0531	0.9545	0.0525	0.9547
		Third order	0.0458	0.9468	0.0454	0.9475
30	20	Jeffreys	0.0450	0.9510	0.0464	0.9494
		Ghosh & Kim	0.0492	0.9473	0.0491	0.9469
		slr	0.0560	0.9467	0.0570	0.9462
		Welch	0.0494	0.9472	0.0492	0.9468
		Third order	0.0529	0.9504	0.0533	0.9507

TABLE 1b: For $\mu_1 = 2.0$, $\mu_2 = 0.0$, $\sigma_2^2 = 1.0$ without loss of generality and for various $n_1 \leq n_2$, σ_1^2, we record the proportion of the 10,000 cases where the true δ is less than the lower limit and less than the upper limit of the 90% central interval.

			$\sigma_1^2 = 2.0$		$\sigma_1^2 = 4.0$	
n_1	n_2	Method	<Lower limit	<Upper limit	<Lower limit	<Upper limit
		Target	0.0500	0.9500	0.0500	0.9500
		Sim SD	0.0022	0.0022	0.0022	0.0022
2	3	Jeffreys	0.0144	0.9836	0.0186	0.9766
		Ghosh & Kim	0.0351	0.9624	0.0438	0.9530
		slr	0.1267	0.8732	0.1363	0.8602
		Welch	0.0460	0.9530	0.0549	0.9440
		Third order	0.0424	0.9586	0.0522	0.9519
2	20	Jeffreys	0.0417	0.9592	0.0456	0.9564
		Ghosh & Kim	0.0969	0.8855	0.0921	0.8930
		slr	0.1479	0.8495	0.1544	0.8429
		Welch	0.0793	0.9232	0.0760	0.9283
		Third order	0.0651	0.9334	0.0640	0.9337
5	7	Jeffreys	0.0375	0.9606	0.0409	0.9587
		Ghosh & Kim	0.0495	0.9504	0.0497	0.9493
		slr	0.0760	0.9257	0.0799	0.9242
		Welch	0.0511	0.9496	0.0507	0.9484
		Third order	0.0518	0.9484	0.0520	0.9488
10	15	Jeffreys	0.0457	0.9540	0.0480	0.9538
		Ghosh & Kim	0.0520	0.9477	0.0535	0.9484
		slr	0.0630	0.9367	0.0630	0.9374
		Welch	0.0518	0.9484	0.0531	0.9485
		Third order	0.0517	0.9481	0.0516	0.9469
15	20	Jeffreys	0.0466	0.9542	0.0485	0.9515
		Ghosh & Kim	0.0509	0.9490	0.0512	0.9477
		slr	0.0577	0.9421	0.0597	0.9410
		Welch	0.0507	0.9492	0.0510	0.9482
		Third order	0.0509	0.9493	0.0516	0.9491
20	30	Jeffreys	0.0549	0.9442	0.0545	0.9445
		Ghosh & Kim	0.0471	0.9533	0.0483	0.9537
		slr	0.0550	0.9443	0.0546	0.9446
		Welch	0.0513	0.9509	0.0496	0.9510
		Third order	0.0491	0.9486	0.0490	0.9504

We find that the third order frequentist procedure gives values that are very close to the target values 0.05 and 0.95, although pushing the three standard deviation simulation limits in the asymmetric sample size cases, with (2, 20) being the most conspicuous.

The Welch procedure, finely turned to the specifics of the Student type quantity (6), comes quite close and indeed does better in several very small sample cases such as (2, 20). Of course with a first sample of size 2 the usual t-statistic for the corresponding mean has a Cauchy distribution and it is known that the third order approach does give a bearable approximation to the Cauchy but does not duplicate an exact Cauchy calculation; see Rekkas et al. (2008).

6. BOX AND COX EXAMPLE

Now consider the Box & Cox (1964) model $y = (X\beta + \sigma z)^{1/\lambda}$, where the error vector z is a sample from a given error distribution. For ease of exposition we use the simple regression version with $y_i = (\alpha + \beta x_i + \sigma z_i)^{1/\lambda}$ and take the errors to be standard normal. The higher-order likelihood procedure however extends to nonnormal error and to nonlinear regression (Fraser, Wong & Wu, 1999) with minor increase in the computational burden, mostly in calculating the needed constrained maximum likelihood values.

Chen, Lockhart & Stephens (2002) questioned the focus on the regression parameter β in Box & Cox (1964) and in Bickel & Doksum (1981), and chose β/σ as having stability for estimation analysis. We extend the questioning concern to both parameters and view them as being artefacts of the notation used and thus not parameters of the original investigation; this is in accord with McCullagh (2002). Correspondingly we do not examine them here directly or in simulations.

The higher-order likelihood methodology allows us to examine almost any smooth parameter and in particular allows us to focus easily on various parameters of direct physical interest in an original regression context. For this we consider the example discussed by Chen, Lockhart & Stephens (2002) which involves gasoline x in litres added to a vehicle and the corresponding distance y in kilometres driven until empty. Chen, Lockhart & Stephens (2002) present data involving 107 values of distance driven in kilometres y_i and of corresponding amount of fuel consumed x_i in litres. We compare the p-value functions obtained by the signed log-likelihood ratio and the third order methodology given as in (3). The Bayesian methodology is not examined because we do not find a satisfactory route to a default prior: the power parameter λ on its own has group properties, and the parameters α, β, and σ on their own have group properties; but the full parameter does not have such properties which seem needed to produce an appropriate default prior. Also the marginalization issues (Dawid, Stone & Zidek, 1973) indicate that it could be difficult to obtain individual priors to target the various component parameters of interest. This is partly an issue of parameter curvature which will be addressed elsewhere.

For an input of some particular amount of gasoline of interest say x_0, we might well be interested in the mean distance traveled

$$\psi(\theta) = E(\alpha + \beta x_0 + \sigma z)^{1/\lambda}.$$

This would seemingly not be functionally accessible given the power transformations and the mean value calculation. Accordingly we follow Yang (2002) and focus directly on the parameter recording the median distance for some fuel input value x_0 of interest; this would have the form

$$\psi_1(\theta) = (\alpha + \beta x_0)^{1/\lambda},$$

which is a composite of α, β, λ, and the fuel input x_0.

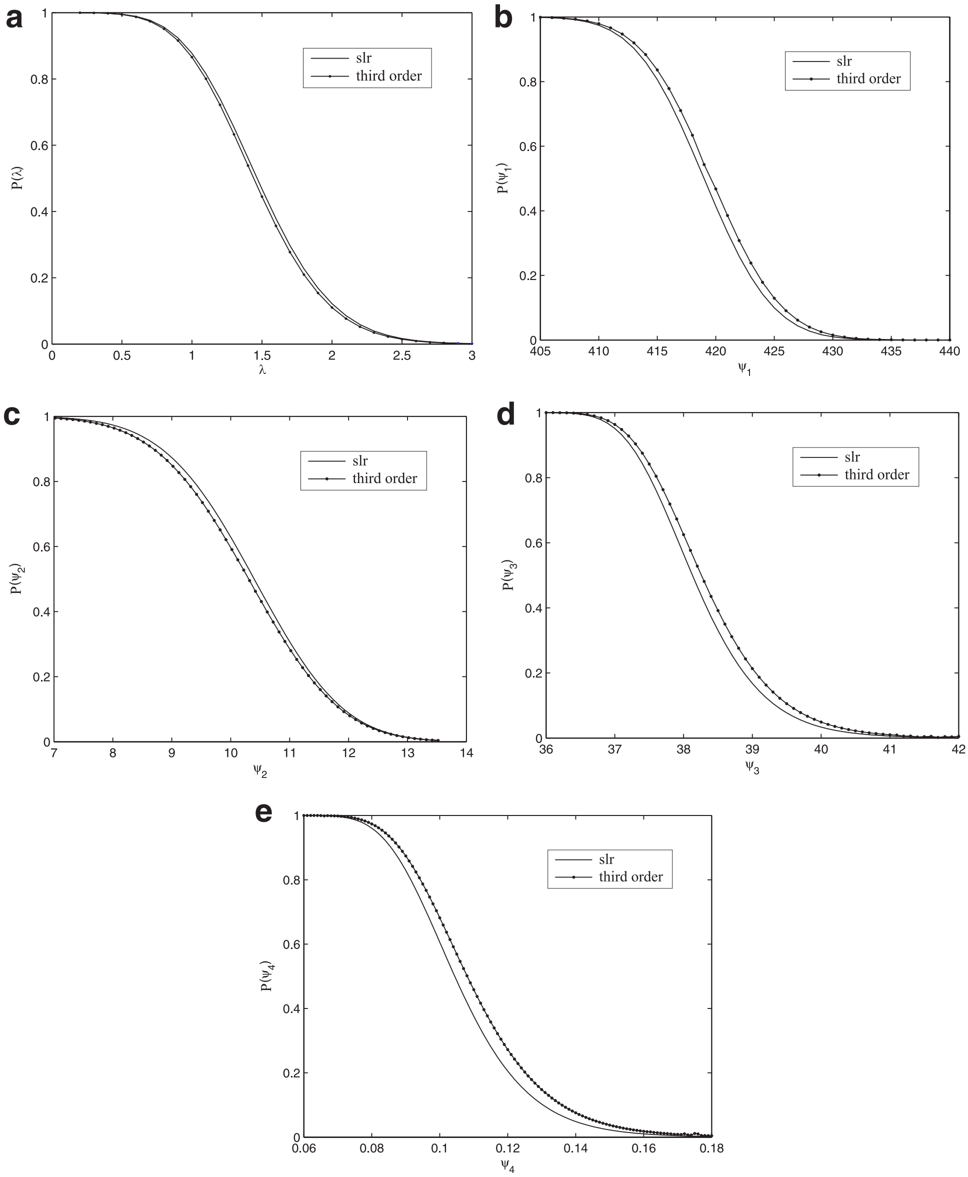

FIGURE 1: The p-value functions for (a) the median distance p(ψ_1) from 30 liters of fuel; (b) the extra distance when topping up 30 liters p(ψ_2); (c) the liters needed to go a median 500 kilometers p(ψ_3); (d) the rate of fuel consumption for a distance of 500 kilometers p(ψ_4); (e) the power parameter λ.

For the data in Chen, Lockhart & Stephens (2002) we plot in Figure 1a the p-value function $p(\psi_1)$ for an input of $x_0 = 30$ L of gasoline; the actual parameter being examined is then $\psi_1 = (\alpha + 30\beta)^{1/\lambda}$, and the 95% confidence interval for ψ_1 in kilometres is

slr:	(409.96, 428.22)
third order:	(410.40, 429.03).

We do note that the data set is very large with $n = 107$ and that likelihood alone is powerful.

To give some indication of the computation needed for the third order, we record some steps indicating what is needed beyond just likelihood. Of course the preceding suggests we do not need more than likelihood in the presence of the large data set from the literature, but the simulations in the next section will focus on performance with smaller data sets.

The calculations for the preceding use an observed log-likelihood, $\ell(\theta)$, and a nominal canonical parameter $\varphi(\theta)$; these are sums, $\ell(\theta) = \sum \ell_i(\theta)$ and $\varphi(\theta) = \sum \varphi_i(\theta)$, of contributions from individual coordinates. In particular, we have

$$\ell_i(\theta) = -\log(\sigma) - \frac{1}{2\sigma^2}(y_i^\lambda - \alpha - \beta x_i)^2 + \log\lambda + (\lambda - 1)\log y_i$$

for the component observed log-likelihood, and

$$\varphi_i(\theta) = \left\{ \frac{-\lambda y_i^{\lambda-1}}{\sigma^2}(y_i^\lambda - \alpha - \beta x_i) + (\lambda - 1) y_i^{-1} \right\} (v_{i1}, v_{i2}, v_{i3}, v_{i4})$$

for the component canonical parameter. The row vector $V_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ is obtained by first differentiating $y = (\alpha + \beta x + \sigma z)^{1/\lambda}$ for fixed z with respect to $(\alpha, \beta, \sigma, \lambda)$, and then evaluating the result at the observed y_i with the overall maximum likelihood value $(\hat\alpha, \hat\beta, \hat\sigma, \hat\lambda)$ and with the corresponding values of the residuals; this gives with observed values:

$$V_i = \frac{y_i}{\hat\lambda y_i^{\hat\lambda}}(1, x_i, \hat z_i, -y_i^{\hat\lambda}\log y_i)$$

and

$$\hat z_i = \hat\sigma^{-1}(y_i^{\hat\lambda} - \hat\alpha - \hat\beta x_i).$$

These are routine problem-specific calculations that more generally can be mechanized by replacing derivatives by high accuracy differences.

But there are other physically meaningful parameters that could also be of interest in an original context, and are also readily accessible from the cumulant generating approach to likelihood. For example, we could be interested in the rate at which extra distance is achieved by providing additional fuel to some initial amount x_0. The interest parameter is then the derivative $(\mathrm{d}/\mathrm{d}x)(\alpha + \beta x)^{1/\lambda}|_{x=x_0}$:

$$\psi_2(\theta) = \beta\lambda^{-1}(\alpha + \beta x_0)^{1/\lambda - 1}.$$

For the data set in Chen, Lockhart & Stephens (2002), we plot in Figure 1b the p-value function $p(\psi_2)$ for the rate of extra distance when topping up 30 L of gasoline; the parameter being examined is then $\psi_2(\theta) = \beta\lambda^{-1}(\alpha + 30\beta)^{1/\lambda - 1}$. The 95% confidence interval for ψ_2 in kilometres per litre is then

slr:	(7.965, 12.709)
third order:	(7.805, 12.682).

Or in a reverse way we could be interested in the amount of fuel needed to go a median distance of say y_0 kilometres. The interest parameter is then the solution for x from the equation $(\alpha + \beta x)^{1/\lambda} = y_0$:

$$\psi_3(\theta) = \beta^{-1}(y_0^\lambda - \alpha).$$

For the large data set, we plot in Figure 1c the p-value function $p(\psi_3)$ for a target distance of $y_0 = 500$ km. The parameter being examined is then $\psi_3(\theta) = \beta^{-1}(500^{\lambda} - \alpha)$. The 95% confidence interval for ψ_3 in litres is then

slr:	(36.83, 40.18)
third order:	(36.90, 40.43).

Or we could be interested in the rate of fuel consumption at a certain terminal distance y_0 of travel; this is closely related to $\psi_2(\theta)$ and is given as

$$\psi_4(\theta) = \beta^{-1}\lambda y_0^{\lambda-1}.$$

For the large data set, we plot in Figure 1d the p-value function $p(\psi_4)$ for a target distance of $y_0 = 500$ km; the interest parameter is then the rate of fuel consumption $\psi_4(\theta) = \beta^{-1}\lambda 500^{\lambda-1}$ for a terminal distance of 500 km. The 95% confidence interval for ψ_4 in litres kilometre is then

slr:	(0.0777, 0.1486)
third order:	(0.0797, 0.1557).

Also in Figure 1e, we plot the p-value function for the power parameter λ obtained from the given data. The 95% confidence interval for λ is

slr:	(0.7126, 2.4100)
third order:	(0.6939, 2.3773).

The plots show that the third order likelihood p-value can differ substantially from the first order approach. In the next section, we report on simulation results that strongly support the third order p-value approach. Then in Appendix, we illustrate with simple examples the flexibility of the higher-order approach.

In concluding this section we do note that the computations sometimes need special attention with presently available likelihood programs. The main computational concern centres on the need for highly accurate and reliable maximum likelihood sub-calculations. For example, the maximum likelihood estimate given in Chen, Lockhart & Stephens (2002) is $\hat{\lambda} = 1.46$, with $\hat{\theta}^o = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}) = (-592.2767, 244.2613, 703.987)$ which yields $\ell(\hat{\theta}^o) = -414.940034$. On the other hand, if we take a more accurate evaluation $\hat{\lambda} = 1.4655$, we obtain $\hat{\theta}^o = (-634.7085, 253.2915, 730.7206)$ which gives $\ell(\hat{\theta}^o) = -414.939951$. We thus obtain a large change in the estimated parameters from a very small change in $\hat{\lambda}$; the observed log-likelihood function however does not seem to be seriously affected by such accuracy, but negative values can occur for $\ell(\hat{\theta}^o) - \ell(\hat{\theta}^o_{\psi})$, and the observed information matrix can become computationally non-positive definite; in other words, the computational accuracy can be very important, particularly when, as here, the parameter effects are closely linked or confounded.

7. BOX AND COX SIMULATIONS

The preceding large sample example with $n = 107$ suggests that with large data sets there may not be a large change in going from first order likelihood to third order likelihood. For simulation in this section, we examine medium sized data sets with $n = 30$. For this we note that the power parameter λ itself is rather special in that such transformations form a group. Accordingly, it suffices to examine a single λ value and, for such, an appropriate power transformation can convert that value to $\lambda = 1$. For the other parameters, we choose $\sigma = 1$, $\alpha = 3$, and examine three choices for $\beta = 1/3$, 1, and 3. For input values x, we look at a moderate range $(x_1, \dots, x_{30}) = (11, \dots, 40)$; the model is

$$y_i = (\alpha + \beta x_i + \sigma z_i)^{1/\lambda}, \quad i = 1, \dots, 30$$

with normally distributed errors z_i.

We generated $N = 10{,}000$ samples of size 30 from the standard normal and then used the model and the particular parameter values to generate 10,000 data sets for each parameter combination. We then calculated the corresponding central 95% confidence intervals for λ and for $\psi_1(\theta)$ based on the signed log-likelihood ratio and the third order method. Table 2 for λ and Table 3

TABLE 2: Box–Cox problem: Interest in λ.

β	Method	Lower limits	Between lower and upper limits	Upper limits
	Target	0.0250	0.9500	0.0250
	Sim SD	0.0016	0.0022	0.0016
1/3	slr	0.0271	0.9396	0.0333
	third order	0.0258	0.9479	0.0263
1	slr	0.0333	0.9334	0.0333
	third order	0.0275	0.9465	0.0260
3	slr	0.0354	0.9310	0.0330
	third order	0.0282	0.9464	0.0254

Simulation coverage probabilities for the 95% confidence interval for λ with $N = 10{,}000$. For $\lambda = 1$ without loss of generality, and with $\sigma = 1$, $\alpha = 3$, $(x_1, \dots, x_{30}) = (11, 12, \dots, 40)$ and various β values, we record the proportion of cases of 95% confidence intervals having the true λ less than the interval, in the interval or above the interval; we also record the simulation standard deviation.

TABLE 3: Box–Cox problem: Interest in $\psi_1(\theta)$.

β	Method	Lower limits	Between lower and upper limits	Upper limits
	Target	0.025	0.9500	0.0250
	Sim SD	0.0016	0.0022	0.0016
1/3	slr	0.0260	0.9364	0.0376
	third order	0.0235	0.9494	0.0271
1	slr	0.0280	0.9352	0.0368
	third order	0.0230	0.9496	0.0274
3	slr	0.0296	0.9354	0.0350
	third order	0.0238	0.9494	0.0268

Simulation coverage probabilities for the 95% confidence interval for $\psi_1(\theta)$ with $N = 10{,}000$. For $\lambda = 1$ without loss of generality, and with $\sigma = 1$, $\alpha = 3$, $(x_1, \dots, x_{30}) = (11, 12, \dots, 40)$ and various β values, we record the proportion of cases of 95% confidence intervals having the true $\psi_1(\theta)$ less than the interval, in the interval or above the interval; we also record the simulation standard deviation.

for $\psi_1(\theta)$ record the proportion of samples where the true value of the parameter falls below the lower confidence limit, within the confidence interval, and above the upper confidence limit. Note that the nominal values are 0.025, 0.95, and 0.025, respectively; simulation standard deviations are reported.

We did not complete simulations for the remaining parameters ψ_2, ψ_3, and ψ_4 but in Section 8 cite various broadly based simulation studies available in the literature.

8. DISCUSSION

Three simple statistical models involving normal distributions have appeared sporadically in the literature over the better part of a century and many statistical inference procedures have been brought to bear with little agreement as to the appropriate direction. We have surveyed these enigmatic examples: the methods that have been applied, some additional methods that are available, and some simulations to guide the choice of method.

The oldest problem, the Behrens (1929)–Fisher (1935a,b) problem, has received various Bayesian and frequentist analyses with little unanimity as to the most inclusive or accurate. We have compared by simulations various Bayesian and likelihood methods and find more reliable intervals using higher-order likelihood method. This is further supported by an $N = 100{,}000{,}000$ McMC simulation for a particular minimum sample size case in Bédard et al. (2008).

A somewhat more recent problem, the Welch (1939)–Cox (1958) problem has received primarily frequentist and decision theoretic attention, but we note that a Bayes (1763) approach addresses the major difficulties almost immediately, as does the higher-order likelihood approach. Much of the controversy in the literature has been concerned with whether to condition or not, and for this simple example, the conditioning is immediate for both of the approaches.

A still more recent problem, the Box & Cox (1964) problem has received primarily frequentist analysis. Much attention has been given to the choice of parameter to analyze and we defer to the view that the parameter chosen should have meaning in the original context for the problem. The absence of Bayesian approaches is partly explained by the difficulty in finding targeted priors appropriate to the various parameters of interest. This seems not an issue from the frequentist view, although more complicated parameters may require more demanding calculations. A partial Bayesian step could involve the elimination of the complicating power parameter by the use of a flat non-informative prior. Such a step has much appeal but some cautions for this have recently appeared in the literature; see Stainforth et al. (2007) and Heinrich (2006).

We have used simulations as needed for the two more recent problems, and obtained support for the use of the higher-order likelihood approach. Also various simulations in the literature are cited at the end of this section. We have focussed on the flexibility of the methods in managing complications in the model. A different issue raised in the editorial process focused on whether the higher-order methods are robust. Our emphasis on obtaining accurate results under variations in the model type is somewhat counter to this, and generally it is found that the results do depend on details of the statistical model: that if you incorporate additional information concerning the form of the model then you can obtain more accurate and precise inference results.

For background, examples, and simulations, see Davision, Fraser & Reid (2006), DiCiccio, Field & Fraser (1990), Fraser (1990, 1991), Fraser & Reid (1993), Fraser, Reid & Wong (1991), Fraser, Reid & Wu (1999), and Fraser, Wong & Wu (1999).

APPENDIX

(i) *The linearized interest parameter*. The rotated coordinate $\chi(\theta)$ in the $\varphi(\theta)$ parameterization is obtained from the gradient vector of $\psi(\theta)$ at $\hat{\theta}^o_\psi$ and has the form

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}^o_\psi)}{|\psi_{\varphi'}(\hat{\theta}^o_\psi)|} \cdot \varphi(\theta).$$

The first factor is the unit row vector version of the gradient vector $\psi_{\varphi'}(\hat{\theta}^o_\psi)$; it is obtained from

$$\psi_{\varphi'}(\theta) = \frac{\partial\psi(\theta)}{\partial\varphi'} = \left(\frac{\partial\psi(\theta)}{\partial\theta'}\right) \cdot \left(\frac{\partial\varphi(\theta)}{\partial\theta'}\right)^{-1} = \psi_{\theta'}(\theta)\varphi_{\theta'}^{-1}(\theta);$$

in this we take $\psi_{\varphi'}$ to be the Jacobian of the column vector ψ with respect to the row vector φ' ; the unit vector is evaluated at the observed data point.

(ii) *Information determinants*. The full information determinant $j_{\theta\theta'}(\hat{\theta}^o)$ contains the negative second derivative of log-likelihood at the maximum, and the corresponding information in the new parameterization is available as

$$|j_{\varphi\varphi'}(\hat{\theta}^o)| = |j_{\theta\theta'}(\hat{\theta}^o)| \cdot |\varphi_\theta(\hat{\theta}^o)|^{-2}$$

using the Jacobian $\varphi_\theta(\theta) = \partial\varphi(\theta)/\partial\theta'$. The nuisance information determinant in a somewhat similar way takes the form

$$|j_{(\lambda\lambda')}(\hat{\theta}^o_\psi)| = |j_{\lambda\lambda'}(\hat{\theta}^o_\psi)| \cdot |\varphi_{\lambda'}(\hat{\theta}^o_\psi)\varphi'_{\lambda'}(\hat{\theta}^o_\psi)|^{-1} = |j_{\lambda\lambda'}(\hat{\theta}^o_\psi)| \cdot |X'X|^{-1}$$

where the right hand determinant uses $X = \varphi_{\lambda'}(\hat{\theta}^o_\psi)$ and corresponds in the regression context to the volume on the regression surface as a proportion of the corresponding volume for regression coefficients; in the preceding formula this changes the scaling for the nuisance parameter to that derived from the φ parameterization. The expressions above are for the case where θ' is given as (ψ, λ'); the more general version without an explicit nuisance parameterization is available in Fraser, Reid & Wu (1999).

(iii) *Examples*

Let y be a sample of size 1 from an exponential distribution with mean θ. For this model, there is no nuisance parameter. The exact p-value function is

$$p(\theta) = 1 - \exp\left\{-\frac{y}{\theta}\right\}.$$

The log-likelihood function and its derivatives are

$$\ell(\theta) = -\log(\theta) - \frac{y}{\theta}, \qquad \ell_\theta(\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}, \qquad \ell_{\theta\theta}(\theta) = \frac{1}{\theta^2} - \frac{2y}{\theta^3};$$

the maximum likelihood value is $\hat{\theta} = y$; and the observed information is $\hat{\jmath} = j_{\theta\theta}(\hat{\theta}) = 1/y^2$. The parameter of interest is taken to be $\psi = \psi(\theta) = \theta$. From (1), the signed log-likelihood ratio is

$$r = r(\psi) = \text{sgn}(y - \psi)\left[2\left\{-\log\frac{y}{\psi} - 1 + \frac{y}{\psi}\right\}\right]^{1/2}.$$

The canonical parameter is $\varphi(\theta) = 1/\theta$. Hence $\varphi_\theta(\theta) = -1/\theta^2$ and thus

$$|j_{\varphi\varphi}(\hat{\theta})| = |j_{\theta\theta}(\hat{\theta})||\varphi_\theta(\hat{\theta})|^{-2} = y^2.$$

Moreover, with $\psi_\theta(\theta) = 1$, we have $\chi(\theta) = \varphi(\theta) = 1/\theta$. Therefore, from (4), we have

$$q = q(\psi) = \operatorname{sgn}(y - \psi)\left|\frac{1}{y} - \frac{1}{\psi}\right| y$$

and the p-value function can be approximated by (3).

To illustrate the accuracy, consider $y = 2.25$. The following table reports p-values obtained from the signed log-likelihood ratio (slr), third order as in (3), and the exact for selected values of ψ.

	ψ				
Method	0.5	1	3	5	10
slr	0.9971	0.8256	0.3918	0.2404	0.1156
third order	0.9887	0.8934	0.5265	0.3621	0.2022
exact	0.9889	0.8946	0.5276	0.3624	0.2015

	ψ				
Method	20	30	50	100	200
slr	0.0536	0.0340	0.0191	0.0088	0.0041
third order	0.1074	0.0732	0.0448	0.0227	0.0115
exact	0.1064	0.0723	0.0440	0.0222	0.0112

Consider another example which has a nuisance parameter. Let $(y_1, \dots, y_n)$ be a sample from a Gamma distribution with shape β and mean μ. Let $t_1 = \sum \log y_i$ and $t_2 = \sum y_i$. Then the log-likelihood function can be written as

$$\ell(\theta) = \ell(\beta, \mu) = -n \log \Gamma(\beta) + n\beta \log \beta - n\beta \log \mu + \beta t_1 - \frac{\beta t_2}{\mu}.$$

The overall maximum likelihood estimate of θ is $\hat{\theta} = (\hat{\beta}, \hat{\mu})'$ where $\hat{\mu} = t_2/n$ and $\hat{\beta}$ satisfies

$$-g(\hat{\beta}) + \log \hat{\beta} - \log\left(\frac{t_2}{n}\right) + \frac{t_1}{n} = 0,$$

where $g(\cdot)$ is the digamma function. Note that $\hat{\beta}$ has to be obtained numerically. By differentiating the log-likelihood function twice, we have

$$j_{\theta\theta'}(\theta) = \begin{pmatrix} ng'(\beta) - n/\beta & n/\mu - t_2/\mu^2 \\ n/\mu - t_2/\mu^2 & -n\beta/\mu^2 + 2\beta t_2/\mu^3 \end{pmatrix}$$

where $g'(\cdot)$ is the trigamma function.

The canonical parameter is $\varphi(\theta) = (\beta, \beta/\mu)'$. Hence

$$\varphi_{\theta'}(\theta) = \begin{pmatrix} 1 & 0 \\ 1/\mu & -\beta/\mu^2 \end{pmatrix}, \qquad \varphi_{\theta'}^{-1}(\theta) = \begin{pmatrix} 1 & 0 \\ \mu/\beta & -\mu^2/\beta \end{pmatrix}$$

and $\varphi_\lambda(\theta) = (0, -\beta/\mu^2)$. The parameter of interest is taken to be the shape parameter β, thus $\psi = \psi(\theta) = \beta$. We then have $\psi_{\theta'}(\theta) = (1, 0)$ and the constrained maximum likelihood estimate of θ is $\hat{\theta}_\psi = (\beta, \hat{\mu})$. Since $\psi_{\varphi'}(\theta) = \psi_{\theta'}(\theta)\varphi_{\theta'}^{-1}(\theta) = (1, 0)$, we have $\chi(\theta) = \beta$. Thus

$$\frac{|j_{\varphi\varphi'}(\hat{\theta})|}{|j_{(\lambda\lambda')}(\hat{\theta}_\psi)|} = \{ng'(\hat{\beta}) - n/\hat{\beta}\}\beta/\hat{\beta}.$$

From (1), we have

$$r = \text{sgn}(\hat{\beta} - \beta)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}$$

and from (4), we have

$$q = \text{sgn}(\hat{\beta} - \beta)|\hat{\beta} - \beta|[\{ng'(\hat{\beta}) - n/\hat{\beta}\}\beta/\hat{\beta}]^{1/2}.$$

Finally, we have $r^* = r - 1/r\ \log(r/q)$.

Wong & Wu (1998) examined the above problem and from their simulations the approximation based on the exact conditional likelihood function gave the best coverage, which we use for comparison with the present third order method.

As the numerical illustration, we examine the Gross & Clark (1975) data in Grice & Bain (1980) giving survival times for 20 mice exposed to 240 rad of gamma radiation.

152	152	115	109	137	88	94	77	160	165
125	40	128	123	136	101	62	153	83	69

The following table reports the p-value obtained by the log-likelihood ratio, the present third order method, and the approximation discussed in Wong & Wu (1998), for various β values. The third order approaches are substantially in agreement and the signed likelihood ratio analysis differs noticeably.

	β						
Method	3	5	7	10	12	15	20
slr	0.9985	0.9527	0.7614	0.3370	0.1464	0.0300	0.0011
third order	0.9963	0.9181	0.67100	0.2492	0.0965	0.0169	0.0005
Wong & Wu	0.9962	0.9174	0.66700	0.2481	0.0959	0.1068	0.0005

ACKNOWLEDGEMENTS

The Natural Sciences and Engineering Research Council of Canada has provided support for this research. We express sincere thanks to the Editor, Associate Editor and referees for very helpful comments and corrections. And very special thanks to Jean-Francois Plante for support with the Abstract.

BIBLIOGRAPHY

T. Bayes (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.

O. E. Barndorff-Nielsen (1986). Inference on full or partial parameters based on the standardized, signed log likelihood ratio. *Biometrika*, 73, 307–322.

O. E. Barndorff-Nielsen & D. R. Cox (1979). Edgeworth and saddle-point approximations with statistical applications (with discussions). *Journal of the Royal Statistical Society B*, 41, 279–312.

O. E. Barndorff-Nielsen & D. R. Cox (1994). *Inference and Asymptotics*, Chapman and Hall, London.

M. Bédard, D. A. S. Fraser & A. Wong (2008). Higher accuracy for Bayesian and frequentist inference: large sample theory for small sample likelihood. *Statistical Science*, 22, 301–321.

W. V. Behrens (1929). Ein Beitrag zur Fehlerbereichnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahresberichte*, 68, 807–837.

P. J. Bickel & K. A. Doksum (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296–311.

G. E. P. Box & D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211–252.

G. E. P. Box & D. R. Cox (1982). An analysis to transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209–210.

G. Chen, R. A. Lockhart & M. Stephens (2002). Box-Cox transformations in linear models: large sample theory and tests of normality. *Canadian Journal of Statistics*, 30, 177–234.

D. R. Cox (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29, 357–372.

H. E. Daniels (1954). Saddlepoint approximation in statistics. *Annals of Mathematical Statistics*, 29, 631–650.

A. Davison, D. A. S. Fraser & N. Reid (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society B*, 68, 495–508.

A. Dawid, M. Stone & J. Zidek (1973). Marginalization paradoxes in Bayesian and structural inference. marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society B*, 35, 189–233.

T. J. DiCiccio, C. A. Field & D. A. S. Fraser (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika*, 77, 77–95.

R. A. Fisher (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, 222, 309–368.

R. A. Fisher (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.

R. A. Fisher (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 23, 528–535.

R. A. Fisher (1934). Two new properties of mathematical likelihood. *Proceedings of the Cambridge Philosophical Society A*, 144, 285–307.

R. A. Fisher (1935a). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98, 39–54.

R. A. Fisher (1935b). The fiducial argument in statistical inference. *Eugenics*, 11, 141–172.

D. A. S. Fraser (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77, 65–76.

D. A. S. Fraser (1991). Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86, 258–265.

D. A. S. Fraser (2004). Ancillaries and conditional inference (with discussion). *Statistical Science*, 19, 333–369.

D. A. S. Fraser & N. Reid (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statistics Sinica*, 3, 67–82.

D. A. S. Fraser & N. Reid (1995). Ancillaries and third order significance. *Utilitas Mathematica*, 47, 33–53.

D. A. S. Fraser & N. Reid (2001). Ancillary information for statistical inference. *Empirical Bayes and Likelihood Inference*, S. E. Ahmed & N. Reid, editors, Springer, New York, pp. 185–207.

D. A. S. Fraser, N. Reid & A. Wong (1991). Exponential linear models: A two pass procedure for saddle point approximation. *Journal of the Royal Statistical Society B*, 53, 483–492.

D. A. S. Fraser, N. Reid & J. Wu (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, 86, 249–264.

D. A. S. Fraser, A. Wong & J. Wu (1999). Regression analysis, nonlinear or nonnormal: Simple and accurate p-values from likelihood analysis. *Journal of the American Statistical Association*, 94, 1286–1295.

M. Ghosh & Y. H. Kim (2001). The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis. *Canadian Journal of Statistics*, 29, 5–17.

J. V. Grice & L. J. Bain (1980). Inference concerning the mean of the gamma distribution. *Journal of the American Statistical Association*, 75, 929–933.

A. J. Gross & V. A. Clark (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*, Wiley, New York.

J. Heinrich (2006). The Bayesian approach to setting limits: what to avoid. *Statistical Problems in Particle Physics, Astrophysics and Cosmology: Proceedings of PHYSTAT05*, L. Lyons & M. Ünel, editors, World Scientific, London, pp. 98–102.

H. Jeffreys (1961). *Theory of Probability*, Oxford University Press, Oxford.

R. Lugannani & S. Rice (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Advance Applied Probability*, 12, 475–490.

P. McCullagh (2002). What is a statistical model? (with discussion). *Annals of Statistics*, 30, 1225–1310.

J. Neyman (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A*, 237, 333–380.

M. Rekkas, Y. She, Y. Sun & A. Wong (2008). An interesting application of a likelihood-based asymptotic. *Journal of Applied Probability and Statistics*, 3, 275–285.

D. A. Stainforth, M. R. Allen, E. R. Tredger & L. A. Smith (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Transactions of the Royal Society A*, 365, 2145–2161. For comment, see also Economist (Aug 18) 2007, p. 69.

B. L. Welch (1939). On confidence limits and sufficiency with particular reference to parameters of location. *Annals of Mathematical Statistics*, 10, 58–69.

B. L. Welch (1947). The generalization of Student's problem when several different populations are involved. *Biometrika*, 34, 28–35.

A. C. M. Wong & J. Wu (1998). Comparisons of approximate tail probabilities for the shape parameter of the gamma distribution. *Computational Statistics and Data Analysis*, 27, 333–344.

Z. Yang (2002). Median estimation through a regression transformation. *Canadian Journal of Statistics*, 30, 235–242.

Received 6 April 2007
Accepted 18 December 2008