The formal relationship between analytic and bootstrap approaches to parametric inference

T.J. DiCiccio

Cornell University, Ithaca, NY 14853, U.S.A.

T.A. Kuffner*

Washington University in St. Louis, St. Louis, MO 63130, U.S.A.

G.A. Young*

Imperial College London, London SW7 2AZ, U.K.

Abstract

Two routes most commonly proposed for accurate inference on a scalar interest parameter in the presence of a (possibly high-dimensional) nuisance parameter are parametric simulation ('bootstrap') methods, and analytic procedures based on normal approximation to adjusted forms of the signed root likelihood ratio statistic. Under some null hypothesis of interest, both methods yield $p-$values which are uniformly distributed to error of third-order in the available sample size. But, given a specific dataset, what is the formal relationship between $p-$values calculated by the two approaches? We show that the two methodologies give the same inference to second order in general: the analytic $p-$value calculated from a dataset will agree with the bootstrap $p-$value constructed from that same dataset to $O(n^{-1})$, where n is the sample size. In practice, the agreement is often startling.

Keywords: $p-$value, Bootstrap, Likelihood, Signed root likelihood ratio statistic, Normal pivot

*Corresponding author

Email addresses: tjd9@cornell.edu (T.J. DiCiccio), kuffner@wustl.edu (T.A. Kuffner), alastair.young@imperial.ac.uk (G.A. Young)

1. Introduction

We are concerned with inference, primarily using the signed root likelihood ratio statistic R, on a scalar interest parameter ψ, in the presence of a (possibly high-dimensional) nuisance parameter ϕ, based on a random sample of size n from an assumed parametric distribution depending on $\theta = (\psi, \phi)$. Two routes most commonly proposed for accurate inference on ψ are parametric simulation ('bootstrap') methods, [see 5, 8] and analytic procedures based on normal approximation to adjusted forms of R, obtained via small-sample asymptotics. Prominent among analytic procedures is use of a normal approximation to the R^* statistic introduced by Barndorff-Nielsen [1, 2]. Our purpose here is to elucidate the formal relationship between the bootstrap approach to inference, specifically as applied to the signed root statistic R, and the analytic approach based on R^*. In this paper, we examine the specific relationships between the bootstrap and analytic methods for estimation of $p-$values for inference on ψ: particular focus in our numerical illustrations will be with estimation of $p-$values under the null hypothesis. We use results from DiCiccio et al. [3, 4] to show that from a theoretical perspective, quite generally, analytic and bootstrap $p-$values are equivalent to $O(n^{-1})$: the two $p-$values constructed from the same dataset agree to that order. Several examples showing close empirical agreement of $p-$values, even for very small sample sizes n, are provided.

2. Problem setting

Suppose $Y = (Y_1, \ldots, Y_n)$ is a continuous random vector whose distribution depends on a parameter $\theta = (\theta^1, \ldots, \theta^d) = (\psi, \phi)$, where ψ is a scalar parameter of interest and ϕ is a vector of nuisance parameters, of dimension $d - 1$. Further suppose that it is required to test the null hypothesis $H_0 : \psi = \psi_0$ against a one-sided alternative. We wish to compare, for a given dataset, the $p-$values derived from analytic approximation to the distribution of R^* with the $p-$values derived from the bootstrap distribution of R.

For testing the null hypothesis against one-sided alternatives, we may use the signed root of the usual likelihood ratio statistic

$$R(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)[2\{L(\hat{\theta}) - L(\hat{\theta}_\psi)\}]^{1/2} = \mathrm{sgn}(\hat{\psi} - \psi)[2\{M(\hat{\psi}) - M(\psi)\}]^{1/2},$$

where $L(\theta)$ is the log-likelihood function, $\hat{\theta} = (\hat{\psi}, \hat{\phi})$ is the global maximum likelihood estimator, $\hat{\theta}_\psi = (\psi, \hat{\phi}_\psi)$ is the constrained maximum likelihood estimator given ψ, and $M(\psi) = L(\hat{\theta}_\psi)$ is the log profile likelihood function for ψ. Under the null hypothesis, the repeated sampling distribution of R is standard normal to error of order $O_p(n^{-1/2})$. The analytic route to achieve higher-order accuracy in such an inferential setting is based on a standard normal approximation to an adjusted version of $R(\psi)$, denoted by $R^*(\psi)$.

The development of $R^*(\psi)$ is as follows. Suppose that the log-likelihood function is written as $L(\theta; \hat{\theta}, a)$, with $(\hat{\theta}, a)$ minimal sufficient, where a is ancillary, having a distribution which, at least approximately, does not depend on θ. Such a decomposition holds, trivially, in full exponential families, where $\hat{\theta}$ is minimal sufficient and no ancillary a is required, and in transformation models, where the maximal invariant serves as the ancillary a. As noted by Severini [13, §6.5], beyond the exponential family and transformation model contexts, it may be difficult to establish that such a decomposition holds, but general approximations, in particular constructions of approximate ancillaries, are possible which still allow validity of the properties discussed here for analytic methods of inference. A drawback of such constructions is that explicit expression of the log-likelihood in terms of $(\hat{\theta}, a)$ may then be intractable. This does not affect the calculation of a bootstrap $p-$value, but would require approximation to the R^* statistic, which we now describe.

The R^* statistic is defined [1, 2] as

$$R^*(\psi) = R(\psi) + R(\psi)^{-1} \log(U(\psi)/R(\psi)),$$

with

$$U(\psi) = \begin{vmatrix} L_{;\hat{\theta}}(\hat{\theta}) - L_{;\hat{\theta}}(\hat{\theta}_\psi) \\ L_{\phi;\hat{\theta}}(\hat{\theta}_\psi) \end{vmatrix} / \{|j_{\phi\phi}(\hat{\theta}_\psi)|^{1/2} |j(\hat{\theta})|^{1/2}\}.$$

Here $j(\theta) = (-L_{rs}(\theta))$ denotes the observed information matrix, with $L_{rs}(\theta) = \partial^2 L(\theta)/\partial\theta^r \partial\theta^s$, where the indices r, s range from $1, \ldots, d$, and $j_{\phi\phi}$ denotes the $(d -$

$1) \times (d-1)$ sub-matrix corresponding to components of the nuisance parameter ϕ. Also,

$$L_{;\hat{\theta}}(\theta) \equiv L_{;\hat{\theta}}(\theta; \hat{\theta}, a) = \frac{\partial}{\partial \hat{\theta}} L(\theta; \hat{\theta}, a), \quad L_{\phi;\hat{\theta}}(\theta) \equiv L_{\phi;\hat{\theta}}(\theta; \hat{\theta}, a) = \frac{\partial^2}{\partial \phi \partial \hat{\theta}} L(\theta; \hat{\theta}, a).$$

The conditional distribution of the test statistic $R^*(\psi) = R + R^{-1} \log(U/R)$ given a, and hence the unconditional distribution under repeated sampling, is standard normal to error of order $O_p(n^{-3/2})$. An alternative to the standard normal distribution for approximating tail probabilities of $R^*(\psi)$ is the generalized Lugannani-Rice formula [2]; to error of order $O(n^{-3/2})$,

$$\mathrm{pr}(R^* \leq r^* \big| a; \theta) = \Phi(r^*) = \Phi(r) + \varphi(r)(1/r - 1/u), \tag{1}$$

where $r^* = r + r^{-1} \log(u/r)$. The simulation route to inference in this setting is based on the parametric bootstrap approximation to the marginal distribution of $R(\psi)$. This is defined as the sampling distribution of $R(\psi)$ under the model specified by parameter value $\hat{\theta}_\psi$, the constrained maximum likelihood estimator for the observed data sample: see DiCiccio et al. [5], Lee & Young [8]. This parametric bootstrap yields $p-$values which are, under repeated sampling and supposing ψ is the true value of the interest parameter, uniformly distributed to error of order $O(n^{-3/2})$.

We consider first a motivating example.

Example 1. *Extreme value location-scale*. Let $\{X_1, \ldots, X_n\}$ be a random sample from the Weibull density

$$f(x; \beta, \gamma) = \gamma \beta (\gamma x)^{\beta - 1} \exp\{-(\gamma x)^\beta\}, \; x > 0,$$

with interest parameter β. Defining $Y_i = \log X_i$, the Y_i are a random sample from an extreme value distribution $EV(\mu, \psi)$, a location-scale family, with scale and location parameters $\psi = \beta^{-1}, \; \mu = -\log \gamma$. Interest is in inference on the scale parameter of the extreme value distribution. This distribution constitutes an ancillary statistic model: inference for ψ conditions on the observed data value of the ancillary $a = (a_1, \ldots, a_n)$, with $a_i = (y_i - \hat{\mu})/\hat{\psi}$. Exact conditional inference is analytically straightforward, but requires numerical integration for its calculation: see, for instance, Pace & Salvan [9, §7.6]. Here, it is easily verified that the conditional distribution of $R(\psi)$ given

a does not depend on the nuisance parameter μ, so the exact conditional inference is equivalent to a 'conditional bootstrap', which would be based on simulating the conditional distribution of $R(\psi)$ given a, modulo the error introduced by the finite simulation required in practice. It is of interest to see how well this exact conditional inference is approximated by a marginal bootstrap, which ignores the conditioning and is based on simulation of the marginal distribution of $R(\psi)$.

Consider the following specific data sample of size $n = 5$, representing the failure times of a set of pressure vessels, as given by Keating et al. [6]: `274, 1661, 1787, 28·5, 236`. We model the data by the Weibull distribution and consider inference on the associated scale parameter ψ in the derived extreme value location-scale model. For a range of values of ψ_0, consider testing $H_0 : \psi = \psi_0$ versus $H_1 : \psi < \psi_0$. We calculate $p-$values obtained by: (i) normal approximation to the distribution of $R(\psi_0)$; (ii) normal approximation to the distribution of $R^*(\psi_0)$; (iii) (marginal) bootstrap approximation to the distribution of $R(\psi_0)$, computed using $B = 500,000$ simulated samples for each ψ_0; (iv) exact conditional inference. In each case, the $p-$value considered as a function of ψ_0 is calculated: these 'significance functions' are shown in Fig. 1. It is clear that $p-$values calculated by normal approximation to the sampling distribution of $R^*(\psi)$ are, for each value of ψ_0 tested, very close to those calculated by a marginal bootstrap, and both are rather indistinguishable from exact conditional inference.

Fig. 2 compares $p-$values calculated by normal approximation to $R^*(\psi)$ and the marginal bootstrap for inference on ψ in the extreme value location-scale model with parameter values $(\mu, \psi) = (0, 1)$, for sample size $n = 5$. For a series of 200 simulated samples the null $p-$values, that is $p-$values for testing the true null hypothesis $\psi = 1$, from the two approaches are plotted against each other, demonstrating clearly their closeness, for all 200 replications.

3. Theory

We show in this Section that in general models $p-$values calculated from a data sample by bootstrap estimation of the sampling distribution of $R(\psi)$, or a class of asymptotically equivalent pivots, agree with those calculated from the same data sam-

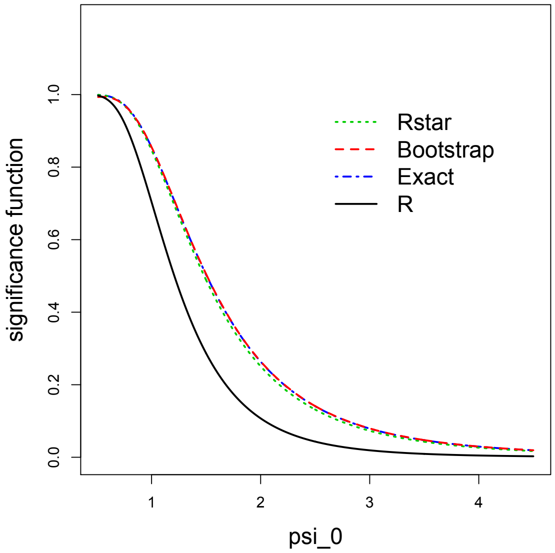

Figure 1: Significance functions, pressure vessel data.

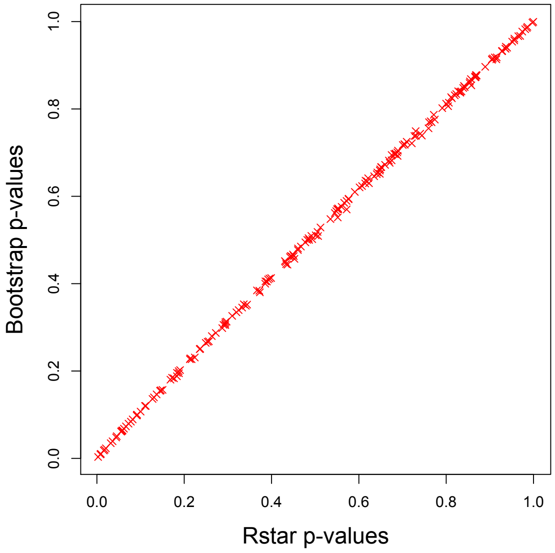

Figure 2: Comparison of null $p-$values, extreme value scale parameter.

ple by normal approximation to the sampling distribution of $R^*(\psi)$ to second-order, $O(n^{-1})$. Both procedures yield $p-$values which, under repeated sampling, are distributed under the null hypothesis as uniform on $(0, 1)$ to error of third-order, $O_p(n^{-3/2})$: see Lee & Young [8]. We provide in the next Section examples demonstrating that, in practice, differences in inferences from the analytic and bootstrap approaches are slight: the analytic $p-$value calculated from a given dataset is typically indistinguishable from the bootstrap $p-$value calculated from that same dataset.

Some further notation is necessary. Arrays and summation are denoted by using the standard conventions, for which the indices $r, s, t, \dots$ are assumed to range over $1, \dots, d$. Summation over the range is implied for any index appearing in an expression both as a subscript and as a superscript. Differentiation is indicated by subscripts, so $L_r(\theta) = \partial L(\theta)/\partial\theta^r$, $L_{rs}(\theta) = \partial^2 L(\theta)/\partial\theta^r\partial\theta^s$, etc. Then $E\{L_r(\theta)\} = 0$; let $\lambda_{rs} = E\{L_{rs}(\theta)\}$, $\lambda_{rst} = E\{L_{rst}(\theta)\}$, etc., and put $l_r = L_r(\theta)$, $l_{rs} = L_{rs}(\theta) - \lambda_{rs}$, $l_{rst} = L_{rst}(\theta) - \lambda_{rst}$, etc. The constants λ_{rs}, $\lambda_{rst}, \dots$, are assumed to be of order $O(n)$. The variables l_r, l_{rs}, l_{rst}, etc., each of which have expectation 0, are assumed to be of order $O_p(n^{1/2})$. The key assumption of our analysis is that joint cumulants of l_r, l_{rs}, etc. are of order $O(n)$. This is a quite standard and weak assumption of regularity in likelihood based inference and is usually satisfied in situations involving independent observations and holds in most models of practical interest, including, for example, regression models: see, for example, Severini [13], Section 3. The analysis of the paper is not, therefore, restrictive: in particular the formal conclusions do not need strong assumptions such as the underlying model being a full exponential family.

We extend the λ-notation by letting $\lambda_{r,s} = E(L_r L_s) = E(l_r l_s)$, $\lambda_{rs,t} = E(L_{rs} L_t) = E(l_{rs} l_t)$, etc. Further, let (λ^{rs}) be the $d \times d$ matrix inverse of (λ_{rs}), and let $\eta = -1/\lambda^{11}$, $\tau^{rs} = \eta\lambda^{1r}\lambda^{1s}$, and $\nu^{rs} = \lambda^{rs} + \tau^{rs}$. Thus, λ^{rs}, τ^{rs}, and ν^{rs} are of order $O(n^{-1})$, while η is of order $O(n)$. For clarity, we point out that a subscript or superscript of '1' refers to the scalar interest parameter ψ, where ψ is the first component of θ.

DiCiccio et al. [4] consider hypothesis testing for ψ based on a test statistic, $T(\psi)$, expressible as $T(\psi) = \eta^{1/2}(T_1 + T_2) + O_p(n^{-1})$, where $T_1 = -\lambda^{1r} l_r$ and T_2 is of the form $T_2 = \xi^{rst} l_{rs} l_t - \xi^{rs} l_r l_s$, with constants ξ^{rst} and ξ^{rs} assumed to be of order

$O(n^{-2})$. This construction includes all commonly used likelihood-based test statistics (DiCiccio et al. [4]); in particular it includes $R(\psi)$. They show that the first three cumulants of $T(\psi)$ are

$$\kappa_1 = E\{T(\psi)\} = \eta^{1/2}\{\xi^{rst}\lambda_{rs,t} + \xi^{rs}\lambda_{rs}\} + O(n^{-1}),$$

$$\kappa_2 = \mathrm{var}\{T(\psi)\} = 1 + O(n^{-1}),$$

$$\kappa_3 = \mathrm{skew}\{T(\psi)\} = \eta^{3/2}(\lambda^{1r}\lambda^{1s}\lambda^{1t}\lambda_{rst} + 3\lambda^{1r}\lambda^{1s}\lambda^{1t}\lambda_{rs,t} - 6\xi^{rs1}\lambda^{1t}\lambda_{rs,t} - 6\xi^{11}) + O(n^{-1}),$$

while the fourth- and higher-order cumulants are of order $O(n^{-1})$ or smaller. For $R(\psi)$, we have $\xi^{rst} = \lambda^{1r}\lambda^{st} + \frac{1}{2}\lambda^{1r}\tau^{st}$ and $\xi^{rs} = \frac{1}{2}\lambda^{1t}\lambda^{ru}\nu^{sv}\lambda_{tuv} + \frac{1}{6}\lambda^{1t}\tau^{ru}\tau^{sv}\lambda_{tuv}$.

Expression (3) of Pierce & Bellio [10], generalising Pierce & Peters [11], introduces quantities $\mathrm{NP}(\psi)$ and $\mathrm{INF}(\psi)$, both of order $O_p(n^{-1/2})$, such that $R^*(\psi) = R(\psi) + \mathrm{NP}(\psi) + \mathrm{INF}(\psi)$. Unconditionally $R^*(\psi)$ has the standard normal distribution to error of order $O(n^{-3/2})$. A detailed analysis of this decomposition is given by DiCiccio et al. [3]. They note that $\mathrm{NP}(\psi)$ and $\mathrm{INF}(\psi)$ are of the form $\mathrm{NP}(\psi) = E\{\mathrm{NP}(\psi)\} + O_p(n^{-1})$ and $\mathrm{INF}(\psi) = E\{\mathrm{INF}(\psi)\} + O_p(n^{-1})$, and establish that

$$\begin{aligned} E\{\mathrm{NP}(\psi)\} &= -\eta^{1/2}\lambda^{1r}\nu^{st}(\lambda_{rs,t} + \frac{1}{2}\lambda_{rst}) + O(n^{-1}), \\ E\{\mathrm{INF}(\psi)\} &= \eta^{1/2}\lambda^{1r}\tau^{st}(\frac{1}{2}\lambda_{rs,t} + \frac{1}{6}\lambda_{rst}) + O(n^{-1}). \end{aligned}$$

Consider again a test statistic of the form $T(\psi) = \eta^{1/2}(T_1 + T_2) + O_p(n^{-1})$, with T_1 and T_2 as above. The Cornish-Fisher expansion shows that $T(\psi) - \frac{1}{6}\kappa_3\{T(\psi)\}^2 - \kappa_1 + \frac{1}{6}\kappa_3$ has a sampling distribution which is standard normal to error of order $O(n^{-1})$. We now investigate how $\mathrm{NP}(\psi)$ and $\mathrm{INF}(\psi)$ arise in the normalized version of $R(\psi)$. From the previous formulae we obtain, for $R(\psi)$, or any statistic $T(\psi)$ satisfying the conditions derived by DiCiccio et al. [4] to produce the same $p-$value when calculated for a given dataset as $R(\psi)$ to error of order $O(n^{-1})$,

$$\begin{aligned} -\kappa_1 + \frac{1}{6}\kappa_3 &= -\eta^{1/2}(\lambda^{1r}\nu^{st}\lambda_{rs,t} - \frac{1}{2}\lambda^{1r}\tau^{st}\lambda_{rs,t} - \frac{1}{6}\lambda^{1r}\tau^{st}\lambda_{rst} + \frac{1}{2}\lambda^{1r}\nu^{st}\lambda_{rst}) + O(n^{-1}) \\ &= NP(\psi) + INF(\psi) + O_p(n^{-1}). \end{aligned}$$

The normalized version of $T(\psi)$ is $N\{T(\psi)\} = \Phi^{-1}[F\{T(\psi)\}]$, in terms of the distribution function $F(\cdot)$ of $T(\psi)$. This normalized version of $T(\psi)$ is, to $O_p(n^{-1})$, of the form $T^*(\psi) = T(\psi) - \frac{1}{6}\kappa_3\{T(\psi)\}^2 + NP(\psi) + INF(\psi)$. Note that $N(\cdot)$ is a monotonic function, so exact inference based on the true sampling distribution of $T(\psi)$ is equivalent to that based on the $N(0,1)$ distribution of the normalized statistic $N\{T(\psi)\}$. A simple delta method calculation then shows that a $p-$value calculated from a dataset using the $N(0,1)$ approximation to the distribution of $T^*(\psi)$ is equivalent to order $O(n^{-1})$ to that based on calculation of a $p-$value from the same dataset using the exact sampling distribution of $T(\psi)$. This exact distribution is approximated to error of order $O(n^{-1})$ by the bootstrap, so consequently we see that the bootstrap $p-$value is equivalent, given a particular dataset, to order $O(n^{-1})$ to that based on the pivot $T^*(\psi)$.

In the case of particular interest when $T(\psi)$ is taken as $R(\psi)$, the skewness κ_3 is of order $O(n^{-1})$, so, given a data sample, the statistic values satisfy $T^*(\psi) = R^*(\psi) + O(n^{-1})$. Therefore, the $p-$value based on the bootstrap distribution of $R(\psi)$ is equivalent to order $O(n^{-1})$ to that based on the standard normal approximation to the sampling distribution of $R^*(\psi)$: the bootstrap and analytic $p-$values calculated from the same dataset agree to $O(n^{-1})$ quite generally. Illustration is given in Fig. 2, where the $p-$value calculated by the bootstrap and the analytic $p-$value obtained from $R^*(\psi)$ are almost coincident for all of the simulated datasets.

4. Examples

Example 2. *Extreme value location-scale, continued.* In the extreme value location-scale problem of inference for the scale parameter ψ, the 'ideal' $p-$value p_I is the exact conditional $p-$value. Though our primary motivation is in evaluation of how close analytic $p-$values are to the unconditional bootstrap $p-$values, it is of some interest to examine which approach yields better approximations to the ideal $p-$value p_I. In principle, given a dataset, the analytic approach approximates the ideal inference to third order, $O(n^{-3/2})$, while an unconditional bootstrap yields an error in estimation of p_I of only second order, $O(n^{-1})$: see DiCiccio et al. [4].

We compare, for different sample sizes n, the average absolute percentage relative error of different approximations to the exact conditional $p-$value p_I for testing H_0 :

n	p_{R^*}	p_{LR}	Bootstrap (p_{boot})			Δ_1	Δ_2	Δ_3
			$B = 10^4$	$B = 10^5$	$B = 10^6$			
5	3.49	0.87	1.40	0.72	0.58	3.61	3.50	3.51
10	1.08	0.24	1.37	0.68	0.53	1.69	1.26	1.21
15	0.57	0.11	1.36	0.60	0.42	1.45	0.81	0.70
20	0.36	0.07	1.35	0.54	0.35	1.39	0.64	0.49
25	0.25	0.05	1.33	0.49	0.30	1.34	0.51	0.33

Table 1: Comparison of $p-$values, extreme value scale parameter. Average absolute percentage relative differences between approximate and ideal $p-$values, and average absolute percentage relative differences Δ between analytic and bootstrap $p-$values.

$\psi = 1$ over 10,000 replications, from the extreme value model with $(\mu, \psi) = (0, 1)$. Specifically, we compare: unconditional bootstrap $p-$values p_{boot} based on simulation sizes $B = 10^4, 10^5, 10^6$; $p-$values p_{R^*} obtained by normal approximation to $R^*(\psi)$; $p-$values p_{LR} obtained by the Lugannani-Rice formula (1) for the conditional tail probability of $R^*(\psi)$. We also calculate the average absolute percentage relative difference between the analytic and bootstrap $p-$values, $\Delta = 100|p_{R^*} - p_{boot}|/p_I$. Results are shown in Table 1. Here $\Delta_1, \Delta_2, \Delta_3$ refer to the cases $B = 10^4, 10^5, 10^6$ respectively.

It is clear from Table 1 that the Lugannani-Rice approximation does approximate the ideal $p-$value very closely. How well the unconditional bootstrap approximates p_I depends noticeably on the simulation size B. For large B, the bootstrap is very competitive with normal approximation to $R^*(\psi)$. On the central issue, of comparing analytic and bootstrap $p-$values p_{R^*} and p_{boot}, we see that for moderate sample sizes, say $n > 10$, the average relative difference is less than 1% of the inferentially correct $p-$value, p_I, provided a large simulation size B is adopted in calculation of the bootstrap $p-$value.

Example 3. *Multi-sample exponential model.* Let Y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, q$ be independent, exponential random variables, with Y_{ij} having mean $1/\phi_j$.

The parameter of interest is defined as

$$\psi = q^{-1} \sum_{j=1}^{q} \exp(-\phi_j c_0),$$

where $c_0 > 0$ is a fixed constant, so that $\theta = (\psi, \phi)$, with the nuisance parameter $\phi = (\phi_2, \ldots, \phi_q)$. This example is considered by Sartori et al. [12], and $q\psi$ may be interpreted as the expected number of items failing by c_0 in a parallel system with failures rates $\phi_1, \ldots, \phi_q$. The interest parameter ψ is a nonlinear function of the canonical parameter in a full exponential family model: calculation of the statistic $R^*(\psi)$ is tractable.

For parameter settings $\psi = \psi_0 = 0.6065, \phi_i = 1, i = 2 \ldots, q$, $q = 5$, we compare ideal $p-$values p_I for testing $\psi = \psi_0$ against $\psi > \psi_0$, as obtained by undertaking a massive simulation to construct the sampling distribution of $R(\psi)$ under the true parameter values, with: $p-$values p_{R^*} obtained by normal approximation to distribution of R^*; $p-$values obtained by bootstrapping the marginal distribution of R, for three simulation sizes in evaluation of each bootstrap $p-$value, $B = 10^4, 10^5, 10^6$. We compare the average absolute percentage relative error of the different approximations to the ideal $p-$values p_I over 5000 replications, for a range of sample sizes n, and calculate the average absolute percentage relative difference $\Delta = 100|p_{R^*} - p_{boot}|/p_I$. Results are given in Table 2. As before, $\Delta_1, \Delta_2, \Delta_3$ refer to the cases $B = 10^4, 10^5, 10^6$ respectively.

As in Example 1, we see that Monte Carlo simulation size has a substantial effect on how well p_{boot} approximates p_I. We note that normal approximation to $R^*(\psi)$ gives greater accuracy in approximation of p_I, in particular for small n. However, the bootstrap approximation improves rapidly with increasing n, and, confirming the main point of our analysis, gives $p-$values close to those from analytic approximation. For $n > 10$, the average difference is less than 1% of the inferentially correct $p-$value, p_I.

Example 4. *Curved exponential family model.* Let Y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, q$ be independent normal random variables with means $\mu_j > 0$ and variances $\psi\mu_j^{1/2}$. This model constitutes a curved exponential family. The parameter of interest is ψ, with $\mu_1, \ldots, \mu_q$ treated as nuisance parameters, $\theta = (\psi, \mu_1, \ldots, \mu_q)$. This example is also considered by Sartori et al. [12]. Now calculation of $R^*(\psi)$ is intractable and

n	p_{R^*}	Bootstrap (p_{boot})			Δ_1	Δ_2	Δ_3
		$B = 10^4$	$B = 10^5$	$B = 10^6$			
10	0.55	1.52	0.95	0.88	1.53	1.04	1.00
15	0.29	1.29	0.62	0.49	1.27	0.63	0.54
20	0.20	1.22	0.49	0.33	1.22	0.49	0.36
25	0.15	1.22	0.44	0.26	1.21	0.43	0.27
30	0.12	1.21	0.42	0.21	1.21	0.42	0.22

Table 2: Comparison of $p-$values, multi-sample exponential model. Average absolute percentage relative differences between approximate and ideal $p-$values, and average absolute percentage relative differences Δ between analytic and bootstrap $p-$values.

we utilise an approximation due to Skovgaard [14].

For parameter settings $\psi = 1, \mu_i = i, i = 1, \ldots, q$, $q = 5$, we compare ideal $p-$values p_I for testing $\psi = 2$ against $\psi > 2$, as obtained by a massive simulation of the exact sampling distribution of $R(\psi)$ under the true parameter values, with: $p-$values p_{skov} obtained by normal approximation to the sampling distribution of the Skovgaard approximation to $R^*(\psi)$; $p-$values obtained by bootstrapping the marginal distribution of $R(\psi)$, again for three simulation sizes, $B = 10^4, 10^5, 10^6$. As before, we compare the average absolute percentage relative error of the different approximations to the ideal $p-$values p_I over 5000 replications and examine the average absolute percentage relative difference $\Delta = 100|p_{skov} - p_{boot}|/p_I$. The results are in Table 3. Again, $\Delta_1, \Delta_2, \Delta_3$ refer to the cases $B = 10^4, 10^5, 10^6$ respectively.

Again, the Monte Carlo simulation size has substantial effect on how well p_{boot} approximates p_I, though the results suggest that the bootstrap and Skovgaard's method give very comparable accuracy in approximation of p_I. However, we note that the bootstrap and Skovgaard approximations are on average closer to one another than to the ideal p_I, even for small sample size n, raising the issue perhaps of whether p_I defined in this way is really 'ideal'.

n	p_{skov}	Bootstrap (p_{boot})			Δ_1	Δ_2	Δ_3
		$B = 10^4$	$B = 10^5$	$B = 10^6$			
10	1.29	1.63	1.07	0.97	1.47	0.89	0.83
15	0.72	1.40	0.73	0.60	1.31	0.57	0.44
20	0.49	1.28	0.60	0.43	1.23	0.49	0.30
25	0.37	1.31	0.54	0.35	1.28	0.44	0.23
30	0.30	1.26	0.48	0.30	1.24	0.42	0.19

Table 3: Comparison of $p-$values, curved exponential family model. Average absolute percentage relative differences between approximate and ideal $p-$values, and average absolute percentage relative differences Δ between analytic and bootstrap $p-$values.

5. Discussion

We have examined in this paper the theoretical higher-order agreement between inferences from analytic and simulation approaches to inference on a scalar interest parameter in the presence of a nuisance parameter, comparing inferences made by analytic approximation to the distribution of the modified signed root statistic $R^*(\psi)$ with those obtained from the bootstrap distribution of the unmodified $R(\psi)$. In general, the analytic and bootstrap $p-$values calculated from a given dataset agree to $O(n^{-1})$. We have provided several empirical examples, demonstrating close agreement of the two $p-$values even for very small sample size n, provided a large simulation is employed in construction of the bootstrap $p-$value.

How can we intuitively understand this result? The $R^*(\psi)$-based p-value agrees to third-order with a *conditional* bootstrap p-value, which we could obtain by simulating the conditional distribution of $R(\psi)$ with the nuisance parameter set at its constrained MLE value, and we are conditioning on the data value of an ancillary statistic. In general this only agrees to second-order with the p-value based on simulating the *marginal* distribution of $R(\psi)$, since the marginal and conditional distributions of $R(\psi)$ only agree to that order. DiCiccio et al. [4] demonstrate that $R(\psi)$ is stable to order $O(n^{-1})$, that is the marginal and conditional distributions of this statistic agree to second order.

We have not sought here to evaluate which of the analytic and bootstrap approaches yields most accurate inference, but instead to demonstrate formally that the two approaches will, quite generally, give $p-$values from a given dataset agreeing to high order. In practice, calculation of a bootstrap $p-$value must be based on finite simulation size B. Full evaluation of this approach requires consideration of the effect of using realistic values of B, rather than large values of B used in our illustrations, but such an analysis is beyond the scope of the present paper. Monte Carlo variability introduced by use of finite B need not necessarily reduce accuracy: see Lee & Young [7].

References

[1] BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

[2] BARNDORFF-NIELSEN, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–63.

[3] DICICCIO, T. J., KUFFNER, T. A. & YOUNG, G. A. (2015a). Quantifying nuisance parameter effects via decompositions of asymptotic refinements for likelihood-based statistics. *J. Stat. Plan. Inference* **165**, 1–12.

[4] DICICCIO, T. J., KUFFNER, T. A., YOUNG, G. A. & ZARETZKI, R. (2015b). Stability and uniqueness of p-values for likelihood-based inference. *Statistica Sinica* **25**, 1355–76.

[5] DICICCIO, T. J., MARTIN, M. A. & STERN, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Can. J. Statist.* **29**, 67–76.

[6] KEATING, J. P., GLASER, R. E. & KETCHUM, N. S. (1990) Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics* **32**, 67–82.

[7] LEE, S. M. S. & YOUNG, G. A. (1999). The effect of Monte Carlo approximation on coverage error of double-bootstrap confidence intervals. *J. R. Statist. Soc.* B **61**, 353–66.

[8] Lee, S. M. S. & Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Stat. Prob. Letters* **71**, 143–53.

[9] Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*. Singapore: World Scientific.

[10] Pierce, D. A. & Bellio, R. (2006). Effects of the reference set on frequentist inferences. *Biometrika* **93**, 425–38.

[11] Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. R. Statist. Soc.* B **54**, 701–37.

[12] Sartori, N., Bellio, R., Salvan, A. & Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters. *Biometrika* **86**, 735–42.

[13] Severini, T. A.(2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

[14] Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–65.