Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families

BY THOMAS J. DICICCIO

Department of Social Statistics, Cornell University, Ithaca,

NY 14853, U.S.A.

tjd9@cornell.edu

AND G. ALASTAIR YOUNG

Department of Mathematics, Imperial College London, London,

SW7 2AZ, U.K.

alastair.young@imperial.ac.uk

Summary

Higher-order inference about a scalar parameter in the presence of nuisance parameters can be achieved by bootstrapping, in circumstances where the parameter of interest is a component of the canonical parameter in a full exponential family. The optimal test which is approximated is a conditional one, based on conditioning on the sufficient statistic for the nuisance parameter. A bootstrap procedure which ignores the conditioning

is shown to have desirable conditional properties, in providing third-order relative accuracy in approximation of p-values associated with the optimal test, in both continuous and discrete models. The bootstrap approach is equivalent to third-order to analytical approaches, and is demonstrated in a number of examples to give very accurate approximations even in very small sample sizes.

Some key words: Bootstrap; Conditional test; Full exponential family; Likelihood; Nuisance parameter; Signed root likelihood ratio statistic.

1. Introduction

Suppose that $Y = (Y_1, \ldots, Y_n)$ is a random vector whose distribution depends on a parameter $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ is a vector of nuisance parameters, and suppose that it is required to test the null hypothesis $H_0 : \psi = \psi_0$ against a one-sided alternative of the form $\psi < \psi_0$ or $\psi > \psi_0$. The parametric bootstrap provides a possible approach: see for example Davison & Hinkley (1997, §4.2.3). Indeed, for such a testing problem, higher-order-accurate inference can be achieved by bootstrap procedures. In particular, inference procedures based on estimation of the sampling distribution of an appropriate statistic under the model in which the nuisance parameters are specified as their constrained maximum likelihood values for the given value ψ_0 of the interest parameter offer third-order accuracy, under repeated sampling, quite generally: see DiCiccio et al. (2001) and Lee & Young (2005).

We will be concerned in this paper with bootstrap procedures based on the signed root likelihood ratio statistic,

$$R = R(\psi_0) \equiv \mathrm{sgn}(\widehat{\psi} - \psi_0)[2\{l(\widehat{\theta}) - l(\widehat{\theta}_0)\}]^{1/2},$$

where $l(\theta)$ is the loglikelihood function, $\widehat{\theta} = (\widehat{\psi}, \widehat{\lambda})$ is the maximum likelihood estimator and $\widehat{\theta}_0 = (\psi_0, \widehat{\lambda}_0)$ is the constrained maximum likelihood estimator of θ given H_0. In broad generality, $R(\psi_0)$ has, under the null hypothesis H_0, the standard normal distribution to error of order $O(n^{-1/2})$, so that standard normal approximation produces p-values that are uniformly distributed, under repeated sampling of Y, to error of first-order $O(n^{-1/2})$. This order of error is reduced to third-order $O(n^{-3/2})$ by bootstrapping, using the procedure suggested by DiCiccio et al. (2001). Given ψ_0, the estimator $\widehat{\theta}_0$ is used to construct a bootstrap distribution of $R(\psi_0)$. With $F_R(r;\theta) = \mathrm{pr}(R \leq r;\theta)$ denoting the distribution function of R under θ, the bootstrap distribution function is $F_R(\cdot;\widehat{\theta}_0)$. Evidence against H_0, in favour of, say, $\psi < \psi_0$ is provided by small values of

r, where r denotes the observed data value of $R(\psi_0)$. The appropriate bootstrap p-value is

$$p_R(\psi_0) = F_R(r; \widehat{\theta}_0). \tag{1}$$

Lee & Young (2005) showed that, under repeated sampling, if H_0 is true, the distribution of the bootstrap p-value $F_R(R; \widehat{\theta}_0)$ is uniform to error of third-order. They showed further that the strategy of bootstrapping at the constrained maximum likelihood estimator achieves reduction in error to third-order not only for the signed root likelihood ratio statistic R, but also for other asymptotically normal test statistics such as Wald and score statistics. The conventional approach of bootstrapping at the global maximum likelihood estimator $\widehat{\theta}$, approximating $F_R(r; \theta)$ by $\mathrm{pr}(R \leq r; \widehat{\theta})$, typically achieves a reduction in error to second-order $O(n^{-1})$ only.

However, if the parameter of interest ψ is a component of the canonical parameter in a full exponential family model, the appropriate inference is a conditional one, based on the conditional distribution of the sufficient statistic for the interest parameter given the observed data value of the sufficient statistic for the nuisance parameters. This conditioning furnishes, in particular, an exact similar test with the finite-sample optimality property of being uniformly most powerful among similar tests; see Lehmann & Romano (2005, §4.4). In practice, however, it may be awkward or impossible to construct the exact similar test, as determination of the p-value requires that the conditional distribution of the sufficient statistic should be known.

The primary contribution of this paper is to establish that in this exponential family context the bootstrap p-value (1), constructed without regard to the desired conditioning of the optimal test, nevertheless approximates the p-value of the exact conditional test to relative error of third-order $O(n^{-3/2})$. Thus, in some generality, the bootstrap method outlined above has the same asymptotic properties as saddlepoint methods developed by Skovgaard (1987) and Barndorff-Nielsen (1986) and studied by Jensen (1992).

2. Theoretical development

Let $Y = (Y_1, Y_2)$ be a random variable whose distribution is a full exponential family model having canonical parameter $\theta = (\psi, \lambda)$ and density that we may suppose to be of the form

$$f_Y(y;\theta) = \exp\{\psi y_1 + \lambda^{\mathrm{T}} y_2 - \kappa(\psi,\lambda) - k(y_1,y_2)\}, \qquad y = (y_1, y_2), \tag{2}$$

where y_2 and λ vectors of dimension q. Let $(Y_{1i}, Y_{2i}), i = 1, \ldots, n$, be a random sample from the distribution of Y. The interest parameter ψ is thus a component of the canonical parameter, though we note that, by reparameterization, our arguments may be applied in circumstances where the interest parameter is a linear combination of canonical parameters, or a ratio of canonical parameters. In the latter case the conditioning statistic will depend on ψ_0; see Barndorff-Nielsen & Cox (1994, §2.5). The sufficient statistic is $\bar{Y} = (\bar{Y}_1, \bar{Y}_2)$, where $\bar{Y}_j = \sum Y_{ji}/n$, and the exact conditional test is based on the distribution of $\bar{Y}_1$, given $\bar{Y}_2 = \bar{y}_2$, the observed data value of $\bar{Y}_2$. Conditioning on $\bar{Y}_2 = \bar{y}_2$ eliminates the nuisance parameter λ. Large values of $\bar{Y}_1$ are evidence against H_0 in favour of $\psi > \psi_0$.

When Y_1 is a continuous variable and Y_2 is either a continuous or a lattice variable, Jensen (1995, formulae 5.2.1 and 5.2.5), showed that, for r of order $O(1)$,

$$\mathrm{pr}(R \geq r | \bar{Y}_2 = \bar{y}_2; \psi) = 1 - \Phi(r) + \phi(r)\Big\{\frac{1}{u} - \frac{1}{r} + O(n^{-3/2})\Big\}, \tag{3}$$

where $r = \mathrm{sgn}(\widehat{\psi} - \psi)[2n\{(\widehat{\psi} - \psi)\bar{y}_1 + (\widehat{\lambda} - \widehat{\lambda}_\psi)^{\mathrm{T}}\bar{y}_2 - \kappa(\widehat{\psi},\widehat{\lambda}) + \kappa(\psi,\widehat{\lambda}_\psi)\}]^{1/2}$, $\kappa_\theta(\widehat{\psi},\widehat{\lambda}) = \bar{y}$, $\kappa_\lambda(\psi,\widehat{\lambda}_\psi) = \bar{y}_2$, and $u = n^{1/2}(\widehat{\psi} - \psi)|\kappa_{\theta\theta}(\widehat{\psi},\widehat{\lambda})|^{1/2}|\kappa_{\lambda\lambda}(\psi,\widehat{\lambda}_\psi)|^{-1/2}$. In these expressions, partial differentiation is indicated by subscripts, so that $\kappa_\theta(\theta) = \partial\kappa(\theta)/\partial\theta$, $\kappa_{\theta\theta}(\theta) = \partial^2\kappa(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$, etc., and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and probability density functions, respectively. Furthermore, Jensen (1995, formula 5.2.11) showed that approximation (3) continues to hold when Y_1 is a lattice variable by

replacing u with $u_\ell = n^{1/2}[1 - \exp\{-(\widehat{\psi} - \psi)\}]|\kappa_{\theta\theta}(\widehat{\psi}, \widehat{\lambda})|^{1/2}|\kappa_{\lambda\lambda}(\psi, \widehat{\lambda}_\psi)|^{-1/2}$. Note that, for r of order $O(1)$, both $\Phi(r)$ and $\phi(r)$ are of order $O(1)$, so that (3) may be rewritten as

$$\mathrm{pr}(R \geq r|\bar{Y}_2 = \bar{y}_2; \psi) = \Big\{1 - \Phi(r) + \phi(r)\Big(\frac{1}{u} - \frac{1}{r}\Big)\Big\}\{1 + O(n^{-3/2})\},$$

or

$$\mathrm{pr}(R \geq r|\bar{Y}_2 = \bar{y}_2; \psi) \simeq 1 - \Phi(r) + \phi(r)(u^{-1} - r^{-1}), \tag{4}$$

where the symbol $\simeq$ denotes an approximation with relative error of order $O(n^{-3/2})$.

The goal is to show that $\mathrm{pr}\{R \geq r; (\psi, \widehat{\lambda}_\psi)\} \simeq \mathrm{pr}(R \geq r|\bar{Y}_2 = \bar{y}_2; \psi)$, when r is of order $O(1)$. This result is derived by integrating, when Y_2 is a continuous variable, or summing, when Y_2 is a lattice variable, the right-hand side of (4) with respect to a saddlepoint expansion of the density of $\bar{Y}_2$, to obtain an expansion for the marginal distribution of R, that is, for $\mathrm{pr}(R \geq r; \theta)$. Then, by evaluating this latter expansion at the particular parameter value $(\psi, \widehat{\lambda}_\psi)$, we see that the right-hand side of (4) is recovered, which confirms the desired result.

To simplify the presentation, the case $q = 1$ is considered. Thus, in what follows, Y_2 is taken to be a real-valued random variable and λ is scalar. The demonstration consists of three steps: first we obtain an approximation to the marginal density of $\bar{Y}_2$ by saddlepoint expansion; then we obtain an approximation to the variable $u^{-1} - r^{-1}$ by ordinary Taylor expansion; and finally we combine the results of the previous two steps by Laplace approximation to find an expansion for the expectation of $u^{-1} - r^{-1}$ taken with respect to the marginal distribution of $\bar{Y}_2$.

Step 1. Standard saddlepoint methods, as developed, for example, by Jensen (1995, formula 2.2.4), show that the marginal density of $\bar{Y}_2$ has the expansion

$$f_{\bar{Y}_2}(\bar{y}_2; \theta) = \left(\frac{nh_{(2)}}{2\pi}\right)^{1/2} \exp(-nh)\Big[1 - \frac{1}{n}\Big\{\frac{1}{8}\frac{h_{(4)}}{(h_{(2)})^2} - \frac{1}{6}\frac{(h_{(3)})^2}{(h_{(2)})^3}\Big\} + O(n^{-2})\Big],$$

where $h = h(\bar{y}_2;\theta) = \widehat{\lambda}_\psi \bar{y}_2 - \lambda \bar{y}_2 + \kappa(\psi,\lambda) - \kappa(\psi,\widehat{\lambda}_\psi)$ and $h_{(k)} = \partial^k h/\partial \bar{y}_2^k, k = 1,\ldots,4$.

Differentiation of the equation $\kappa_\lambda(\psi,\widehat{\lambda}_\psi) = \bar{y}_2$, which defines $\widehat{\lambda}_\psi$ as a function of $\bar{y}_2$, yields $\partial\widehat{\lambda}_\psi/\partial\bar{y}_2 = \{\kappa_{\lambda\lambda}(\psi,\widehat{\lambda}_\psi)\}^{-1}$, and hence $h_{(1)} = \widehat{\lambda}_\psi - \lambda$ and $h_{(2)} = \{\kappa_{\lambda\lambda}(\psi,\widehat{\lambda}_\psi)\}^{-1}$. Thus, the point $\bar{y}_2^*$ at which $h(\bar{y}_2;\theta)$ attains its minimum value satisfies $\widehat{\lambda}_\psi = \lambda$, i.e., $\bar{y}_2^* = \kappa_\lambda(\psi,\lambda)$. Let $\delta = \bar{y}_2 - \bar{y}_2^*$, which is of order $O(n^{-1/2})$. Taylor expansion about $\bar{y}_2^*$ in the saddlepoint expansion of the marginal density of $\bar{Y}_2$ yields

$$\begin{aligned} f_{\bar{Y}_2}(\bar{y}_2;\theta) &= \left(\frac{nh^*_{(2)}}{2\pi}\right)^{1/2} \exp(-\tfrac{1}{2}\delta^2 n h^*_{(2)})\{1 + \delta H_1 + \delta^3 n H_3 \\ &\qquad + n^{-1}H_0 + \delta^2 H_2 + \delta^4 n H_4 + \delta^6 n^2 H_6 + O(n^{-3/2})\}, \end{aligned}$$

where $h^*_{(2)} = h_{(2)}(\bar{y}_2^*) = \{\kappa_{\lambda\lambda}(\psi,\lambda)\}^{-1}$ and the coefficients $H_0, H_1, \ldots$ are of order $O(1)$ and are functions of $\kappa_{\lambda\lambda}(\psi,\lambda)$, $\kappa_{\lambda\lambda\lambda}(\psi,\lambda)$, and $\kappa_{\lambda\lambda\lambda\lambda}(\psi,\lambda)$.

Step 2. Standard Taylor expansion of r^2 about $\widehat{\psi}$ yields an expansion of $u^{-1} - r^{-1}$ of the form

$$u^{-1} - r^{-1} = n^{-1/2} g_1(\widehat{\psi},\widehat{\lambda}) + n^{-1} g_2(\widehat{\psi},\widehat{\lambda}) r + O(n^{-3/2}),$$

where $g_1(\psi,\lambda)$ and $g_2(\psi,\lambda)$ are of order $O(1)$ and are functions of the derivatives of $\kappa(\psi,\lambda)$. If we regard $(\widehat{\psi},\widehat{\lambda})$ as a function of $(r,\bar{y}_2)$, this expansion can be used to obtain an expansion for $u^{-1} - r^{-1}$ about $\bar{y}_2^*$, with r set at its specified value. Derivatives relevant to this purpose include

$$\begin{aligned} \frac{\partial\widehat{\psi}(r,\bar{y}_2)}{\partial\bar{y}_2} &= -\frac{(\widehat{\lambda} - \widehat{\lambda}_\psi)}{(\widehat{\psi} - \psi)} \frac{\widehat{\kappa}_{\lambda\lambda}}{|\widehat{\kappa}_{\theta\theta}|} - \frac{\widehat{\kappa}_{\psi\lambda}}{|\widehat{\kappa}_{\theta\theta}|} = -\frac{1}{2}(\widehat{\psi} - \psi)\frac{\widehat{B}}{|\widehat{\kappa}_{\theta\theta}|} + O(n^{-1}), \\ \frac{\partial\widehat{\lambda}(r,\bar{y}_2)}{\partial\bar{y}_2} &= \frac{(\widehat{\lambda} - \widehat{\lambda}_\psi)}{(\widehat{\psi} - \psi)} \frac{\widehat{\kappa}_{\psi\lambda}}{|\widehat{\kappa}_{\theta\theta}|} + \frac{\widehat{\kappa}_{\psi\psi}}{|\widehat{\kappa}_{\theta\theta}|} = \frac{1}{\widehat{\kappa}_{\lambda\lambda}}\Big\{1 + \frac{1}{2}(\widehat{\psi} - \psi)\frac{\widehat{\kappa}_{\psi\lambda}\widehat{B}}{|\widehat{\kappa}_{\theta\theta}|}\Big\} + O(n^{-1}), \end{aligned}$$

where $\widehat{B} = \widehat{\kappa}_{\psi\psi\lambda} - 2\widehat{\kappa}_{\psi\lambda\lambda}\widehat{\kappa}_{\psi\lambda}\widehat{\kappa}_{\lambda\lambda}^{-1} + \widehat{\kappa}_{\lambda\lambda\lambda}\widehat{\kappa}_{\psi\lambda}^2\widehat{\kappa}_{\lambda\lambda}^{-2}$ and $\widehat{\kappa}_{\theta\theta} = \kappa_{\theta\theta}(\widehat{\theta})$, and so on. The quantity $\widehat{B}$ is of order $O(1)$ and arises in the expansion

$$\frac{(\widehat{\lambda} - \widehat{\lambda}_\psi)}{(\widehat{\psi} - \psi)} = \widehat{\kappa}_{\lambda\lambda}^{-1}\{\widehat{\kappa}_{\psi\lambda} - \tfrac{1}{2}(\widehat{\psi} - \psi)\widehat{B}\} + O(n^{-1}).$$

The crucial observation is that the derivatives of $\widehat{\psi}(r, \bar{y}_2)$ and $\widehat{\lambda}(r, \bar{y}_2)$ of all orders taken with respect to $\bar{y}_2$ are of order $O(1)$ or smaller.

It follows that $u^{-1} - r^{-1}$ has an expansion of the form

$$\{u(r, \bar{y}_2)\}^{-1} - r^{-1} = \{u(r, \bar{y}_2^*)\}^{-1} - r^{-1} + \delta n^{-1/2} G^* + O(n^{-3/2}),$$

where G^* is of order $O(1)$ and is a function of r and the derivatives of $\kappa(\psi, \lambda)$ evaluated at $(\widehat{\psi}^*, \widehat{\lambda}^*)$, the value of $(\widehat{\psi}, \widehat{\lambda})$ corresponding to $(r, \bar{y}_2^*)$ with r set at its specified value.

Step 3. The final step is to use Laplace approximation to integrate $\{u(r, \bar{y}_2)\}^{-1} - r^{-1}$ with respect to the marginal density of $\bar{Y}_2$. This integration amounts to finding the expected value of

$$\begin{aligned}
[\{u(r, \bar{y}_2^*)\}^{-1} &- r^{-1} + Dn^{-1/2} G^* + O(n^{-3/2})]\{1 + DH_1 + D^3 nH_3 \\
&\qquad + n^{-1} H_0 + D^2 H_2 + D^4 nH_4 + D^6 n^2 H_6 + O(n^{-3/2})\} \\
&= [\{u(r, \bar{y}_2^*)\}^{-1} - r^{-1}](1 + DH_1 + D^3 nH_3) + Dn^{-1/2} G^* + O(n^{-3/2}),
\end{aligned}$$

where $D = \bar{Y}_2 - \bar{y}_2^*$, and $\bar{Y}_2$ is to be regarded as a normally distributed random variable with mean $\bar{y}_2^*$ and variance $\{nh_{(2)}^*\}^{-1}$. The integration process then yields

$$\{u(r, \bar{y}_2^*)\}^{-1} - r^{-1} + O(n^{-3/2}).$$

If, given $\bar{Y}_2 = \bar{y}_2$, the parameter is taken to be $(\psi, \lambda) = (\psi, \widehat{\lambda}_\psi)$ with $\kappa_\lambda(\psi, \widehat{\lambda}_\psi) = \bar{y}_2$, then $\bar{y}_2^* = \bar{y}_2$, and the average value of $u^{-1} - r^{-1}$ against the marginal density of $\bar{Y}_2$ is $\{u(r, \bar{y}_2)\}^{-1} - r^{-1} + O(n^{-3/2})$, verifying the desired result.

Suppose that Y_2 is a lattice variable whose minimal span is the integers. The density of $\bar{Y}_2$ (Jensen, 1995, formula 2.2.4) has the expansion

$$\begin{aligned}
f_{\bar{Y}_2}(\bar{y}_2; \theta) &= \left(\frac{h_{(2)}^*}{2\pi n}\right)^{1/2} \exp(-\tfrac{1}{2}\delta^2 n h_{(2)}^*)\{1 + \delta H_1 + \delta^3 nH_3 \\
&\qquad + n^{-1} H_0 + \delta^2 H_2 + \delta^4 nH_4 + \delta^6 n^2 H_6 + O(n^{-3/2})\}.
\end{aligned}$$

Thus, averaging $\{u(r,\bar{y}_2)\}^{-1} - r^{-1}$ with respect to the marginal density of $\bar{Y}_2$ requires calculation of a sum of the form

$$\sum_{\delta} (2\pi n)^{-1/2} \exp(-\tfrac{1}{2}\delta^2 n h^*_{(2)}) q(\delta),$$

where $q(\delta) = (h^*_{(2)})^{1/2} \left([\{u(r,\bar{y}^*_2)\}^{-1} - r^{-1}](1 + \delta H_1 + \delta^3 n H_3) + \delta n^{-1/2} G^* + O(n^{-3/2}) \right)$. The discrete version of Laplace approximation detailed by Jensen (1995, formula 3.2.10) shows that the sum equals

$$(h^*_{(2)})^{-1/2} q(0) + \tfrac{1}{2} n^{-1} (h^*_{(2)})^{-3/2} q_{(2)}(0) + O(n^{-2}) = \{u(r,\bar{y}^*_2)\}^{-1} - r^{-1} + O(n^{-3/2}),$$

exactly as in the continuous case, where $q_{(2)}(\delta) = \partial^2 q(\delta)/\partial\delta^2$. Again, provided we set the parameter to be $(\psi,\lambda) = (\psi, \widehat{\lambda}_\psi)$, for which $\kappa_\lambda(\psi, \widehat{\lambda}_\psi) = \bar{y}_2$, the observed value of the conditioning statistic $\bar{Y}_2$, then the sum becomes $\{u(r,\bar{y}_2)\}^{-1} - r^{-1} + O(n^{-3/2})$, as desired.

Suppose finally that Y_1 is a lattice variable. Taylor expansion shows that $u_\ell^{-1} - r^{-1}$ admits an expansion of the same form as for $u^{-1} - r^{-1}$:

$$u_\ell^{-1} - r^{-1} = n^{-1/2} g_{\ell 1}(\widehat{\psi}, \widehat{\lambda}) + n^{-1} g_{\ell 2}(\widehat{\psi}, \widehat{\lambda}) r + O(n^{-3/2}),$$

say. Arguments identical to those used for $u^{-1} - r^{-1}$ apply, and the desired result holds therefore in this case also.

The above argument holds for values of r in the normal deviation region, where R is of order $O(1)$. Jensen (1992) showed that conditional tail probability approximations derived from (4) have relative error of order $O(n^{-1})$ in large deviation regions, where R is of order $O(n^{1/2})$. Consequently, the bootstrap tail probabilities also approximate the conditional ones to the same order of error in those regions.

In the present context of inference about canonical parameters in exponential families, Jensen (1986a) considered bias- and variance-corrected versions of the signed root likelihood ratio statistic R. He showed using an Edgeworth expansion argument that

the conditional distribution of $\{R - \widehat{\mu}(\widehat{\theta}_0)\}/\widehat{\sigma}(\widehat{\theta}_0)$, given $\bar{Y}_2 = \bar{y}_2$, is standard normal to error of order $O(n^{-3/2})$ in the normal deviation region, though uniformity of relative errors in the large deviation region does not hold. Here, Jensen took $\widehat{\mu}(\theta)$ and $\widehat{\sigma}^2(\theta)$ to be asymptotic analytical approximations to the unconditional mean $\mu(\theta)$ and variance $\sigma^2(\theta)$ of R obtained by the delta method. However, $\mu(\widehat{\theta}_0)$ and $\sigma(\widehat{\theta}_0)$ are readily estimated by simulation, so by bootstrapping we may replace the need for analytical bias and variance correction; an alternative bootstrap approach, which only requires estimation of the unconditional mean and variance of R, rather than estimation of the full bootstrap distribution of R, which is likely to be significantly more demanding computationally, approximates the exact conditional p-value by

$$\Phi[\{r - \mu(\widehat{\theta}_0)\}/\sigma(\widehat{\theta}_0)]. \tag{5}$$

Of the two approximate conditional p-values that can be obtained by simulation, $p_R(\psi_0)$, as given by (1), and the p-value (5) obtained by normal approximation, the former will typically be more accurate, particularly in the region of extreme values of r. The primary advantage of (5) is in its computational savings, which could be significant when confidence limits are being constructed.

Approximation (4) furnishes directly an analytical approximation to the $p-$value of the exact conditional test, with the same property of relative error of order $O(n^{-3/2})$ as we have shown for the bootstrap $p-$value (1). An asymptotically equivalent formulation of (4) is

$$\mathrm{pr}(R \geq r|\bar{Y}_2 = \bar{y}_2; \psi) \simeq 1 - \Phi(r^*), \tag{6}$$

where r^* is the adjusted form of the signed root likelihood ratio statistic (Barndorff-Nielsen, 1986) which has the form $r^*(\psi) = r + r^{-1}\log(u/r)$.

In the empirical examples studied in the next section, we are concerned primarily with the accuracy of bootstrap $p-$values (1) and (5) as approximations to those of the

exact conditional test. Recall that the $p-$values (1) are based on the bootstrap sampling distribution of R, while the approximation (5) uses the bootstrap simulation only to derive bias and variance corrections to the statistic. However, we offer also comparisons with analytical approximations to the exact conditional $p-$values derived from (6), as these provide some benchmark of what is achievable by fully analytical means for small sample sizes. In this regard, there is some evidence in the literature which favours use of approximation (6) over (4): see for example Pierce & Peters (1992) and the discussion of Barndorff-Nielsen & Cox (1994, §6.7).

3. Examples

3 · 1. Lognormal mean

Let $Y_1, \ldots, Y_n$ be independent $N(\mu, \sigma^2)$ and suppose we are interested in inference about $\psi = \mu + \frac{1}{2}\sigma^2$, with σ^2 as the nuisance parameter. The inference problem is equivalent to that about the mean of the associated lognormal distribution.

As considered by Jensen (1986a), we assume a data configuration in which the usual unbiased estimator $\tilde{\sigma}^2 = n\widehat{\sigma}^2/(n-1)$ of σ^2 takes the value $\tilde{\sigma}^2 = 5$, with the sample mean $\bar{Y} = -\frac{1}{2}\tilde{\sigma}^2$. With sample size $n = 5$, this gives $\widehat{\psi} = -1/2$. We consider testing $H_0 : \psi = \psi_0$, for a range of values ψ_0, for each value testing against the appropriate one-sided alternative, so that the bootstrap p-value is $\min\{p_R(\psi_0), 1 - p_R(\psi_0)\}$, with $p_R(\psi_0)$ given by (1).

Calculation of the p-values associated with the exact conditional test is awkward, requiring numerical integration, though quite feasible; details are given by Land (1971). In Fig. 1(a) we compare the absolute relative errors of the bootstrap p-values, and also those obtained from the adjusted signed root likelihood ratio statistic $r^*(\psi_0)$ defined by (6), by comparison with the p-values of the exact conditional test. We include in our

comparison both the bootstrap p-values (1) and the p-values (5). All bootstrap p-values were obtained on the basis of 50 million bootstrap samples. Such a large simulation was used in the study to eliminate Monte Carlo variability. We offer some comments in §4 on the size of bootstrap simulation which might more realistically be used in practice.

In the figure, as in those in our other examples, the vertical lines correspond to exact conditional p-values of 0.01. The smallest values of ψ_0 considered in this example correspond to exact conditional p-values as low as 0.0003, and the largest values of ψ_0 considered to p-values around 0.0080. The full bootstrap approximation (1) displays remarkably low relative error, except for extreme left-tailed p-values, and is more accurate than the analytical approximation. The procedure based on normal approximation to a bias and variance corrected version of R constructed by bootstrapping, (5), is remarkably accurate relative to analytical approximation, but is less accurate than the full bootstrap approach for extreme p-values.

Following Davison et al. (2006), we note that, if the relative error of an approximation to the exact conditional p-value is of order $O(n^{-c})$, a graph of the logarithm of the relative error against $\log(n)$ will be linear, with slope $-c$. In Fig. 1(b), we have assumed the same data configuration as before, but varied n, in a test of $H_0 : \psi = 0$. Four approximations to the exact conditional p-value are considered: the bootstrap approximations (1) and (5), the analytical approximation $\Phi(r^*)$, and $\Phi(r)$, obtained by normal approximation to the distribution of the unadjusted signed root statistic. The relative errors show the expected dependence on the sample size n, the lines in the figure demonstrating the $O(n^{-3/2})$ convergence rate of both bootstrap-based approximations and the analytical approximation $\Phi(r^*)$, and the slower rate of convergence $O(n^{-1/2})$ of the approximation $\Phi(r)$. Both bootstrap approximations yield lower relative errors than the analytical method. We note that here, for each sample size n, the exact conditional p-value being approximated is not extreme, being around 0.47, so it is not surprising that the bootstrap

procedure (5) based on normal approximation is competitive with the full bootstrap approximation (1).

3 · 2. Gamma distribution

Now suppose that $Y_1, \ldots, Y_n$ is an independent sample from a gamma distribution with mean μ, shape parameter ν and density

$$f(y; \mu, \nu) = \frac{\nu^\nu}{\Gamma(\nu)} \exp\Big[-\nu\Big\{\frac{y}{\mu} - \log\left(\frac{y}{\mu}\right)\Big\}\Big]\frac{1}{y}, \quad y > 0, \quad \mu, \nu > 0.$$

We consider first inference about the shape parameter ν, with μ nuisance. Here the optimal conditional test is based on the conditional distribution of $\sum_{i=1}^n \log Y_i$ given the observed data value of $\sum_{i=1}^n Y_i$, which is equivalent (Pace & Salvan, 1997, Example 5.14) to testing $H_0 : \nu = \nu_0$ based on a statistic, W say, which is the ratio between the geometric mean and arithmetic mean of the sample observations, in the conditional distribution given the observed value u of $U = \sum_{i=1}^n Y_i$. In fact W and U are independent, so that the conditioning of the optimal test is carried out automatically when referring to the marginal distribution of W. This distribution is complicated (Keating et al., 1990), but easily simulated. We approximated the distribution by generating 200 million samples of the appropriate size n from the gamma density $f(\cdot; 1, \nu_0)$.

We illustrate the bootstrap approximation to the exact conditional p-values on a subset of size $n = 10$ of the dataset considered by Fraser et al. (1997), concerning survival times of mice exposed to gamma radiation. The observations are 152, 115, 109, 152, 137, 88, 94, 77, 160 and 165. As in our previous example we consider testing $H_0 : \nu = \nu_0$ for a range of values of ν_0. The smallest value of ν_0 considered corresponds to an exact conditional p-value of 0.0012, and the largest to an exact $p-$value of 0.0035. Bootstrap p-values are constructed on the basis of 5 million bootstrap samples, and Fig. 2(a) depicts the absolute relative errors for the bootstrap approximations (1) and (5), and for the analytical approximation (6). Some Monte Carlo variability is apparent in

the figure: this is due primarily to the fact that we have simulated the exact conditional p-values, rather than to simulation error in the bootstrap calculation itself. The basic message, however, is clear. Throughout the range of ν_0 considered, which includes very extreme values, the bootstrap approximation (1) displays a small relative error, less than 2%, and compares favourably in terms of accuracy with the analytical method. The simpler bootstrap approximation (5) is again very competitive compared to the analytical method, but is less accurate than the full bootstrap approximation in the extreme tails.

Suppose now that the mean μ is the interest parameter, with the shape parameter ν nuisance. Construction of the optimal conditional test again requires numerical integration of a conditional density, but in this case the conditional density is not explicitly known (Jensen, 1986b), except for sample sizes $n = 2$ or 3. We subject the approximation methods to a stern test, based on the sample of $n = 2$ observations consisting of the first two survival times in the mice dataset considered above. Fig. 2(b) displays the absolute relative errors of the analytical and bootstrap approximations, the latter again based on 5 million samples, to p-values of the optimal conditional test for inference on $H_0 : \mu = \mu_0$. For this extreme dataset the exact p-value is around 0.05 even for values μ_0 in the region of 400, and is only as low as 0.01 for values of μ_0 around 2000. The relative errors are comparatively large, up to around 30%, but again both bootstrap approaches are accurate compared to analytical approximation, with approximation (1) favoured in accuracy terms to (5).

3 · 3. Binomial distribution

Our third example considers a discrete-data model and concerns inference about the difference of log-odds for two binomial distributions. We assume a data configuration as considered by Davison et al. (2006), with Y_1, Y_2 independent binomial random variables, Y_1 being $\mathrm{Bi}(2n, p_1)$ and with log-odds $\lambda = \log\{p_1/(1-p_1)\}$ and Y_2 being $\mathrm{Bi}(2n, p_2)$, with

log-odds $\lambda + \psi = \log\{p_2/(1-p_2)\}$. The observed values are $y_1 = n, y_2 = n + n^{1/2}$ and we are interested in testing no difference in log-odds, $H_0 : \psi = 0$, considering the effect of increasing $n = 4, 9, 16, \ldots$.

An immediate issue arises about the appropriate reference p-value of the optimal conditional test. It is widely argued that the appropriate p-value for inference in such settings is the mid-p-value; see for instance the bibliographical notes of Brazzale et al. (2007, Ch. 3). In the context of our full exponential family model (2), when testing $H_0 : \psi = \psi_0$ against $\psi > \psi_0$, the mid-p-value is

$$\mathrm{pr}(\bar{Y}_1 > \bar{y}_1 | \bar{Y}_2 = \bar{y}_2) + \frac{1}{2}\mathrm{pr}(\bar{Y}_1 = \bar{y}_1 | \bar{Y}_2 = \bar{y}_2).$$

We intend to undertake a full evaluation of the properties of the bootstrap procedures for discrete distributions in subsequent work, but here we provide just illustration that the method provides accurate approximation to the mid-p-value. Figure 3 plots the logarithms of the relative errors of the analytical approximations $\Phi(r)$ and $\Phi(r^*)$, as approximations to the mid-p-value, as well as that of the bootstrap approximations (1) and (5), as n increases. Note that in our analysis the version of r^* used is the continuous form (6) described in §2, rather than a version which incorporates an explicit correction for discreteness. This approach, of using the continuous form of r^* for discrete problems, has been advocated strongly by Pierce & Peters (1999). In this discrete problem, the bootstrap p-values may be obtained by a complete enumeration of the bootstrap distribution, without Monte Carlo sampling. We see from the figure the order $O(n^{-1})$ rate of convergence to the exact conditional mid-p-value of the analytical approximation $\Phi(r^*)$ discussed by Davison et al. (2006) for this example, and discussed in general terms by Brazzale et al. (2007, p. 168). The bootstrap approach based on (1) produces smaller relative errors than the analytical method. In this case, where the exact conditional p-values being approximated are of the order 0.15, the simpler

bootstrap approximation (5) produces relative errors quite indistinguishable from those of the analytical approximation (6).

4. Discussion

From a repeated sampling perspective, the bootstrap provides third-order accuracy in inference about a scalar parameter of interest in the presence of a nuisance parameter quite routinely. We have shown that the same accuracy is obtained in a multi-parameter exponential family context by a bootstrap scheme in which the nuisance parameter is specified as its constrained maximum likelihood value for the null hypothesis value of the interest parameter, the accuracy now being with respect to the p-values of an exact conditional test. The bootstrap procedure displays the same theoretical error properties as analytical, saddlepoint approximation methods, and is conceptually easier since it is applied without regard to the conditioning, though more computationally demanding. In a range of examples studied we have observed the bootstrap approximations to be more accurate than analytical methods.

Conditioning to eliminate nuisance parameters in exponential family models is only one role for conditioning in statistical inference. The other main context, arguably more difficult, concerns conditioning on ancillary statistics in more general models, in particular transformation models. In this context, the unconditional bootstrap procedure can be expected to yield only second-order conditional accuracy in general. Very intricate calculations show the signed root likelihood ratio statistic to be only second-order stable: the marginal distribution of R differs from its distribution conditional on an ancillary statistic at second-order, $O(n^{-1})$. The bootstrap procedure approximates the marginal distribution to order $o(n^{-1})$, but the relevant conditional distribution only to order $O(n^{-1})$.

A fuller analysis of the theoretical behaviour of the bootstrap approach in discrete data settings seems worthwhile. In particular, the rate of approximation of the bootstrap p-value to the exact conditional mid-p-value and its relationship to the approximate conditioning ideas of Pierce & Peters (1999) warrant further analysis.

Finally, we remark briefly on computational considerations. Our numerical illustrations have involved very large bootstrap simulation sizes. In practice a more modest simulation is likely to be desirable. We have carried out extensive investigations into the issue of the size of simulation typically required to reduce the Monte Carlo variability of the bootstrap calculation to a level which allows the good theoretical accuracy properties to be captured in practice. Overall, a bootstrap simulation involving a few tens of thousands of bootstrap samples is seen as advisable. This recommendation is for a somewhat larger simulation than is typically advocated in more straightforward applications of the bootstrap, such as bias estimation or confidence-interval construction, where bootstrap simulation sizes of a few hundreds or thousands are generally advised. An illustration is provided in Fig. 4, concerning the gamma shape parameter inference considered above. Now we consider specifically testing $H_0 : \nu = 30.0$, repeating the bootstrap estimation of the p-value 1000 times, each estimate being based on the simulation of B bootstrap samples, for a range of values of B. Histograms of the absolute relative errors obtained from the 1000 replications of the bootstrap calculation are compared with the error obtained by analytic approximation. Figure 2(a) shows that an 'infinite bootstrap' yields a very small absolute relative error in this situation, substantially smaller than that obtained by the analytic approximation method. However, Fig. 4 shows that only for large bootstrap simulation size $B = 50,000$ does the bootstrap yield, on average, as accurate an estimate of the exact conditional p-value as the analytical approximation (6). For smaller values of B, the relative error of the bootstrap estimator is quite variable. Recall, however, that a very basic property of our assumed model class, a full exponential family, is that

the loglikelihood is strictly concave, ensuring existence and uniqueness of the maximum likelihood estimator. Therefore, though the maximum likelihood estimators required for construction of the bootstrap distribution of the test statistic may not be available in closed form, no numerical problem can arise in implementation. A bootstrap simulation of the size that we have advocated is generally, therefore, quite trivial computationally, within standard computational platforms.

Theoretical considerations similar to those outlined in §2 show that bootstrapping other asymptotically normal test statistics, such as the Wald and score statistics, at the constrained maximum likelihood estimator produces approximations to exact conditional p-values having the same asymptotic properties as those enjoyed by the approximations obtained from bootstrapping R. Nevertheless, in numerical examples, we found that bootstrapping R tends to yield more accurate results, especially in small sample sizes, than does bootstrapping other test statistics, and hence R is the focus of the present paper. However, the loss in accuracy incurred by bootstrapping a different test statistic might be offset by computational advantages, especially in the case of the Wald statistic, since the bootstrap procedure appears to benefit from a large simulation size. This trade-off deserves further study. For the Wald statistic, the choice of parameterization could be a key factor in determining the accuracy achieved for a given simulation size.

REFERENCES

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

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

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge: Cambridge University Press.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press

Davison, A. C., Fraser, D. A. S. & Reid, N. (2006). Improved likelihood inference for discrete data. *J. R. Statist. Soc.* B **68**, 495-508.

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.

Fraser, D. A. S., Reid, N. & Wong, A. (1997). Simple and accurate inference for the mean of the gamma model. *Can. J. Statist.* **25**, 91-9.

Jensen, J. L. (1986a). Similar tests and the standardized log likelihood ratio statistic. *Biometrika* **73**, 567-72.

Jensen, J. L. (1986b). Inference for the mean of a gamma distribution with unknown shape parameter. *Scand. J. Statist.* **13**, 135-51.

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika* **79**, 693-703.

Jensen, J. L. (1995). *Saddlepoint Approximations*. Oxford: Clarendon Press.

Keating, J. P., Glaser, R. E. & Ketchum, N. S. (1990). Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics* **32**, 67-82.

Land, C. E. (1971). Confidence intervals for linear functions of normal mean and variance. *Ann. Statist.* **42**, 1187-205.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses, 3rd ed.* New York: Springer

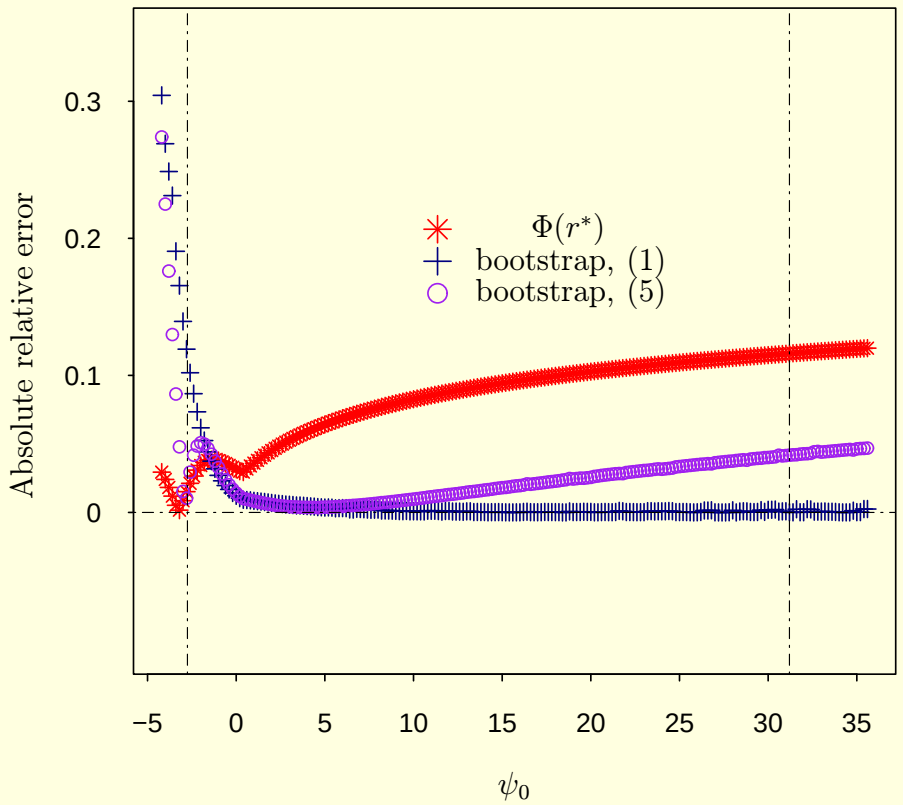
Lee, S. M. S. & Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statist. Prob. Lett.* **71**, 143-53.

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

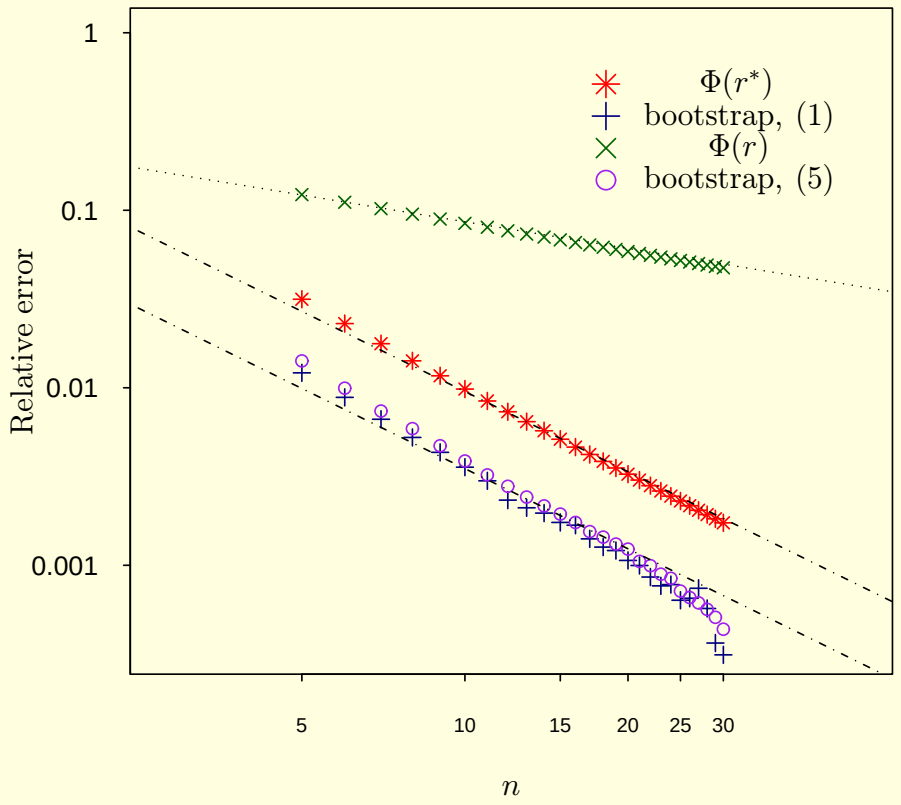
Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with Discussion). *J. R. Statist. Soc.* B **54**, 701-37.

Pierce, D. A. & Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika* **86**, 265-77.

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24**, 875-87.

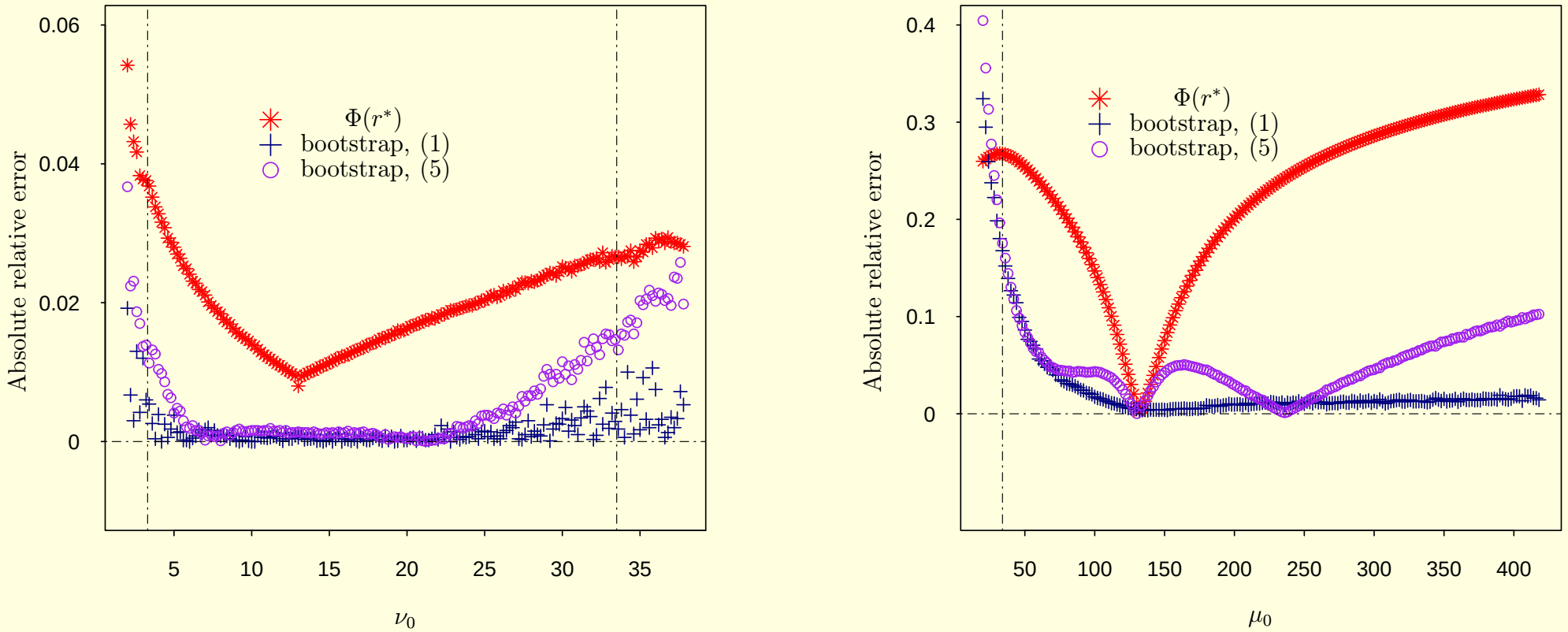


(a)



(b)

Figure 1: Lognormal mean example. (a) Absolute relative errors of approximations of conditional p-values. (b) Logarithmic plot of relative errors of approximations against sample size n.



(a) (b)

Figure 2: Absolute relative errors of approximations of conditional p-values, (a) gamma shape example, (b) gamma mean example.

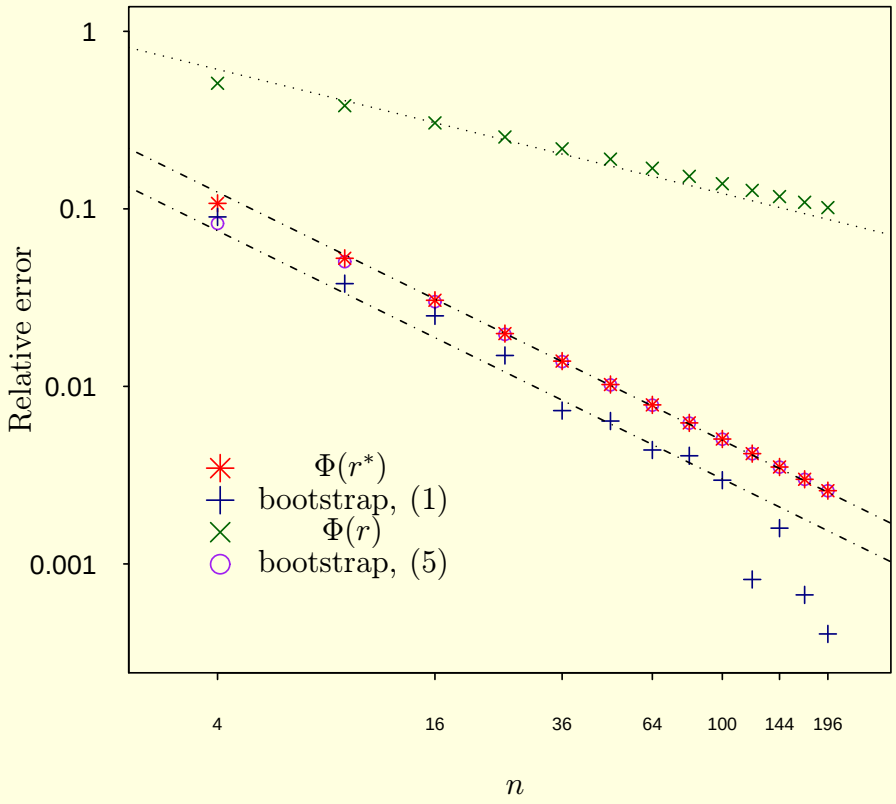


Figure 3: Logarithmic plot of relative errors of approximations of conditional p-values against sample size n, binomial example.

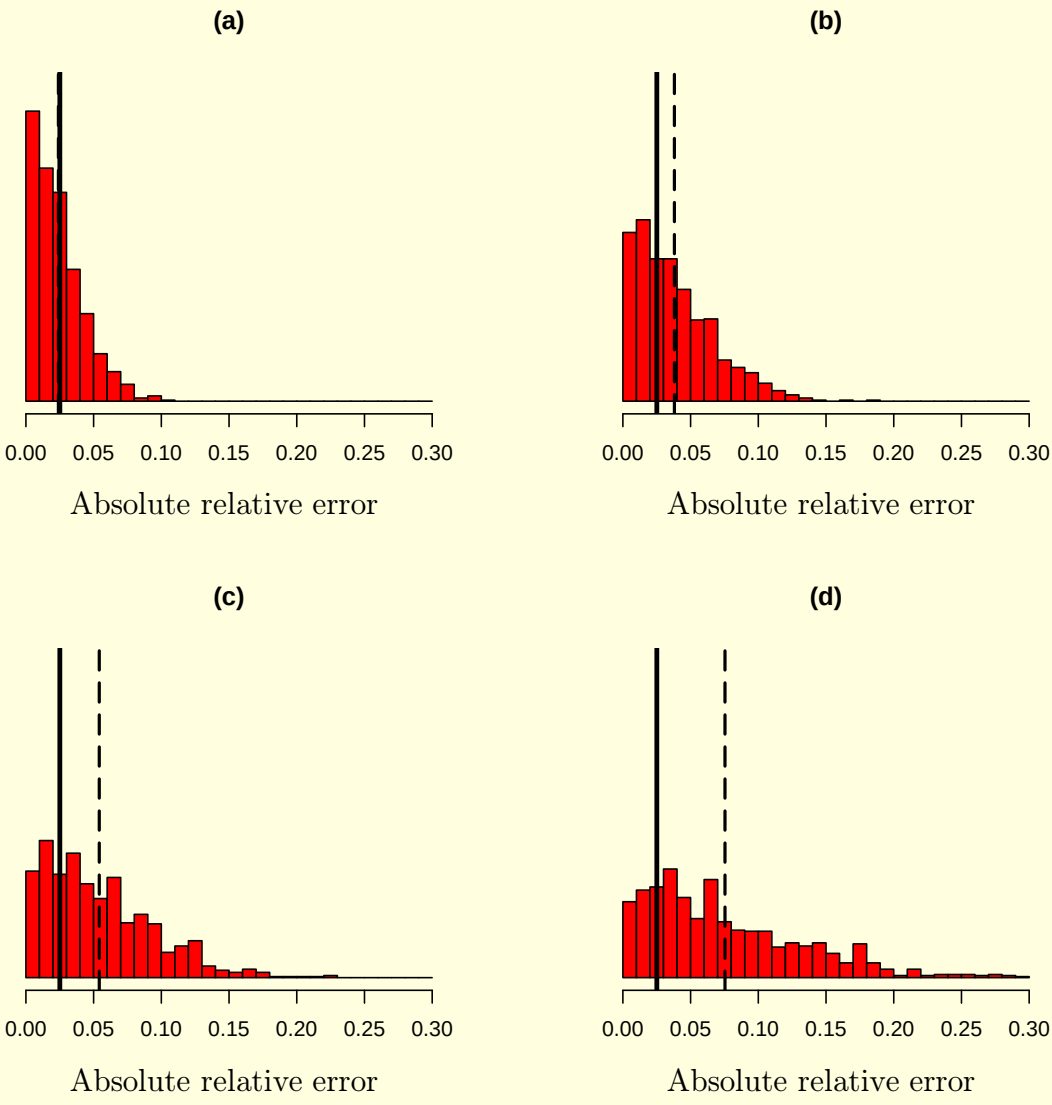


Figure 4: Histogram of absolute relative errors of 1000 bootstrap estimators based on different simulation sizes B, testing gamma shape $\nu = 30$. Solid vertical line indicates absolute relative error of analytic approximation $\Phi(r^*)$, broken vertical line the average of the 1000 bootstrap absolute relative errors. (a) $B = 50000$, (b) $B = 20000$, (c) $B = 10000$, (d) $B = 5000$.