Biometrika (2016), **103**, 3, *pp.* 641–652
Printed in Great Britain

doi: 10.1093/biomet/asw021
Advance Access publication 2 August 2016

Distribution of likelihood-based *p*-values under a local alternative hypothesis

By STEPHEN M. S. LEE

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
smslee@hku.hk

and G. ALASTAIR YOUNG

Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.
alastair.young@imperial.ac.uk

Summary

We consider inference on a scalar parameter of interest in the presence of a nuisance parameter, using a likelihood-based statistic which is asymptotically normally distributed under the null hypothesis. Higher-order expansions are used to compare the repeated sampling distribution, under a general contiguous alternative hypothesis, of p-values calculated from the asymptotic normal approximation to the null sampling distribution of the statistic with the distribution of p-values calculated by bootstrap approximations. The results of comparisons in terms of power of different testing procedures under an alternative hypothesis are closely related to differences under the null hypothesis, specifically the extent to which testing procedures are conservative or liberal under the null. Empirical examples are given which demonstrate that higher-order asymptotic effects may be seen clearly in small-sample contexts.

Some key words: Alternative hypothesis; Asymptotic normality; Bootstrap; Constrained bootstrap; Likelihood; Null hypothesis; p-value; Power; Size.

1. Introduction

Testing of a null hypothesis against a specified alternative by calculation of a p-value is an intrinsic part of statistical inference. Yet it is rare that the sampling distribution of the statistic used for a hypothesis test is known exactly under the null hypothesis in question, typically because of the presence of nuisance parameters that remain unspecified under the hypothesis. Usually, therefore, the test is conducted by calculation of an approximate p-value, either by analytical means or by bootstrap estimation of the null sampling distribution. The sampling distribution of p-values calculated from the exact null sampling distribution of the test statistic in question is, under the null hypothesis, exactly uniform on (0, 1); but in general the null sampling distribution of an approximate p-value is only asymptotically uniform.

A highly useful approach to testing a hypothesis on a parameter of interest in the presence of a nuisance parameter is furnished by procedures based on the likelihood function, including tests based on the likelihood ratio statistic. Although no explicit optimality criteria are invoked, a quite general asymptotic distribution theory allows straightforward implementation of such methods in a wide class of problems.

In this paper we are concerned with inference on a scalar parameter of interest, in the presence of a nuisance parameter, using a likelihood-based statistic which is asymptotically distributed as standard normal, $N(0, 1)$, under a certain null hypothesis. We will focus in particular on comparison of the repeated sampling distribution of p-values calculated from the asymptotic normal approximation to the null sampling distribution of the statistic with the distribution of p-values calculated by bootstrap approximations to the sampling distribution of the statistic (DiCiccio et al., 2001; Lee & Young, 2005; Stern, 2006). In some generality (Lee & Young, 2005), p-values approximated analytically or by bootstrapping are known to be asymptotically uniform under the null hypothesis, with the sampling distribution of p-values obtained by bootstrap approximation being more uniformly distributed under the null hypothesis than those calculated from a normal approximation. However, to discriminate more fully between different p-value approximations, it is necessary to consider also the sampling distribution of p-values when the alternative hypothesis is true. In this paper a higher-order comparison, which generalizes that of Lee & Young (2005) to consider distributions under an alternative hypothesis, is made between p-values obtained by normal approximation and by bootstrap approximation. A key methodological conclusion drawn in this paper is that the results of comparisons of different testing procedures in terms of power under an alternative hypothesis are closely related to differences under the null hypothesis, specifically the extent to which testing procedures are conservative or liberal under the null. This finding provides some validation for the principle of choosing a testing procedure that yields size as close as possible to a nominal desired level, without reference to power under any specified alternative.

2. Inference problem

Suppose that $Y = (Y_1, \ldots, Y_n)$ is a random sample from an unknown underlying distribution F_η indexed by $\eta = (\eta^1, \ldots, \eta^d) \in \mathbb{R}^d$. Let $\theta = g(\eta)$ be a scalar parameter of interest, for some smooth function $g : \mathbb{R}^d \to \mathbb{R}$. Denote by $l(\eta)$ the loglikelihood function based on Y. Let $\hat{\eta} = \arg\max_\eta l(\eta)$ be the global maximum likelihood estimator, and let $\hat{\eta}_\vartheta = \arg\max_\eta \{l(\eta) : g(\eta) = \vartheta\}$ be the constrained maximum likelihood estimator of η for any $\vartheta \in \mathbb{R}$. Typically, we will have $\eta = (\theta, \xi)$, with inference required for the parameter of interest θ in the presence of the nuisance parameter ξ.

Let $\theta_0 = g(\eta_0)$ be a hypothesized value of θ, and suppose that we wish to test the null hypothesis $H_0 : \theta = \theta_0$ against a one-sided alternative, specified as $H_\mathrm{a} : \theta < \theta_0$ or $H_\mathrm{a} : \theta > \theta_0$. The test is performed using a statistic $T(\theta_0) = T(Y, \theta_0)$, and we will assume that large positive values of $T(\theta_0)$ constitute evidence against H_0 in favour of the specified alternative H_a. A key choice for the test statistic $T(\theta_0)$ is based on the signed root likelihood ratio statistic

$$R(\theta_0) = \mathrm{sgn}(\hat{\theta} - \theta_0)\left[2\{l(\hat{\eta}) - l(\hat{\eta}_{\theta_0})\}\right]^{1/2},$$

where $\hat{\theta} = g(\hat{\eta})$. We have that $R(\theta_0)$ is distributed under H_0 as standard normal, with an error of order $n^{-1/2}$. Large positive values of $T(\theta_0) = R(\theta_0)$ are evidence against $H_0 : \theta = \theta_0$ in favour of $H_\mathrm{a} : \theta > \theta_0$, while evidence against H_0 in favour of the alternative $H_\mathrm{a} : \theta < \theta_0$ would be provided by large positive values of $T(\theta_0) = -R(\theta_0)$. In our empirical studies in § 4 we will concentrate on using such a statistic $T(\theta_0)$, which is known (DiCiccio et al., 2015) to have desirable properties compared to other likelihood-based statistics for the inference problem being considered, although other choices of statistic $T(\theta_0)$ are covered by the theory presented in § 3. Examples include the studentized maximum likelihood estimator, or Wald statistic, standardized versions of the profile score, and the signed root of various adjusted forms of the likelihood ratio statistic. For further discussion and references, see Lee & Young (2005, Remark 3).

Assuming an $N(0, 1)$ null distribution for $T(\theta_0)$, the p-value for testing H_0 is approximated by

$$\hat{P}_{\mathrm{N}} = 1 - \Phi\{T(\theta_0)\}.$$

Lee & Young (2005) considered two bootstrap p-values. The constrained bootstrap estimate of the p-value is

$$\hat{P}_{\mathrm{cB}} = 1 - G\big\{T(\theta_0); \hat{\eta}_{\theta_0}, \theta_0\big\}.$$

Here $G(\cdot\,; \hat{\eta}_{\theta_0}, \theta_0)$ denotes the distribution function of $T(Y^*_{\theta_0}, \theta_0)$, where $Y^*_{\theta_0}$ is a random bootstrap sample of n observations drawn from $F_{\hat{\eta}_{\theta_0}}$. This procedure imposes the null hypothesis constraint in the generation of bootstrap samples.

As before, let $\hat{\theta} = g(\hat{\eta})$ be the global maximum likelihood estimator of θ. Denote by Y^* a random bootstrap sample of n observations drawn from $F_{\hat{\eta}}$. The unconstrained parametric bootstrap estimates the null distribution of $T(\theta_0)$ by the bootstrap distribution of $T(Y^*, \hat{\theta})$, and the unconstrained bootstrap estimate of the p-value is therefore defined as

$$\hat{P}_{\mathrm{B}} = 1 - G\big\{T(\theta_0); \hat{\eta}, \hat{\theta}\big\}.$$

This procedure does not impose the null hypothesis constraint in the bootstrapping.

Lee & Young (2005) considered the distribution of the p-values $\hat{P}_{\mathrm{N}}$, $\hat{P}_{\mathrm{cB}}$ and $\hat{P}_{\mathrm{B}}$ under the null hypothesis. They showed that when the statistic $T(\theta_0)$ is distributed as $N(0, 1)$ with an error of $O(n^{-\beta/2})$, the p-values $\hat{P}_{\mathrm{N}}$, $\hat{P}_{\mathrm{cB}}$ and $\hat{P}_{\mathrm{B}}$ are distributed as Un(0, 1), with errors of $O(n^{-\beta/2})$, $O(n^{-(\beta+2)/2})$ and $O(n^{-(\beta+1)/2})$, respectively. Therefore, the discrepancies between the actual and nominal sizes of tests based on the three procedures are of these orders. Considering for illustration the case where the statistic is the signed root likelihood ratio statistic, we have $\beta = 1$, and the normal, constrained bootstrap and unconstrained bootstrap p-values are distributed, when the null hypothesis is true, as uniform with corresponding orders $O(n^{-1/2})$, $O(n^{-3/2})$ and $O(n^{-1})$. Our primary interest here is in examining the distribution of the p-values under an alternative hypothesis. Theoretical results derived in Supplementary Material are summarized in § 3.

3. Main results

Let $\theta_0 = g(\eta_0)$ be a contiguous hypothesized value of θ such that $\delta \equiv \eta - \eta_0 = O(n^{-1/2})$. The usual local alternative formulation has $\delta = O(n^{-1/2})$ but $\delta \neq o(n^{-1/2})$.

Denote derivatives of g and l by $g_i(\eta) = \partial g(\eta)/\partial \eta^i$, $l_i(\eta) = \partial l(\eta)/\partial \eta^i$, $l_{ij}(\eta) = \partial^2 l(\eta)/(\partial \eta^i \partial \eta^j)$, and so on. Write $s(\eta) = n^{-1} l(\eta)$, $s_i(\eta) = n^{-1} l_i(\eta)$, $s_{ij}(\eta) = n^{-1} l_{ij}(\eta)$, and so on. Define $J_{ij}(\eta, \eta_0) = -E_\eta\{s_{ij}(\eta_0)\}$ and $L_{ijk}(\eta) = E_\eta\{s_{ijk}(\eta)\}$. Denote by $J^{ij}(\eta, \eta_0)$ the (i, j)th element of the inverse of the matrix $\{J_{ij}(\eta, \eta_0)\}$. Define $\sigma^2(\eta) = g_i(\eta) g_j(\eta) J^{ij}(\eta, \eta)$, where summation over the range $1, \ldots, d$ is understood for any index appearing once as a subscript and once as a superscript. For brevity we write $\check{f} = f(\eta_0)$ for any function $f(\eta)$ evaluated at $\eta = \eta_0$. Thus we have $\check{s}_i = s_i(\eta_0)$, $\check{g}_i = g_i(\eta_0)$, $\check{J}_{ij} = J_{ij}(\eta_0, \eta_0)$, $\check{\sigma} = \sigma(\eta_0)$, and so on.

Consider a test statistic $T(\theta_0) = T(Y, \theta_0)$ which admits an expansion of the form

$$T(\theta_0) = \pm n^{1/2} \check{\sigma}^{-1} \check{g}_a \check{s}_b \check{J}^{ab} + \Delta_n(\eta, \eta_0), \qquad (1)$$

where $\Delta_n(\eta, \eta_0) = O_{\mathrm{p}}(n^{-1/2})$ can be expanded as a sum of nonrandom multiples, possibly depending on n, of products of quantities $\check{s}_i - E_\eta(\check{s}_i)$, $\check{s}_{ij} - E_\eta(\check{s}_{ij})$, $\check{s}_{ijk} - E_\eta(\check{s}_{ijk}), \ldots$. All commonly used likelihood-based asymptotically normal statistics admit an expansion of this form; see Lee & Young (2005). The sign in (1) is determined by the direction of the one-sided alternative hypothesis H_{a} against which $T(\theta_0)$ serves as a test statistic. Consider testing H_0 in a test of nominal size α, so that the normal approximation, constrained bootstrap and unconstrained bootstrap procedures respectively reject H_0 if $\hat{P}_{\mathrm{N}} < \alpha$, $\hat{P}_{\mathrm{cB}} < \alpha$ and $\hat{P}_{\mathrm{B}} < \alpha$. The discrepancy between the actual and nominal sizes of the test based on the normal approximation is $\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{N}} \leqslant \alpha) - \alpha$, and similarly for the constrained and unconstrained bootstrap procedures. When the statistic $T(\theta_0)$ is distributed under H_0 as $N(0,1)$ with an error of $O(n^{-1/2})$, as is the case for the signed root likelihood ratio statistic, the discrepancies are, respectively, $O(n^{-1/2})$, $O(n^{-3/2})$ and $O(n^{-1})$ for the normal approximation, constrained bootstrap and unconstrained bootstrap tests (Lee & Young, 2005).

In the Supplementary Material, higher-order expansions are derived for the distribution functions of the p-values $\hat{P}_{\mathrm{N}}$, $\hat{P}_{\mathrm{cB}}$ and $\hat{P}_{\mathrm{B}}$. These expansions allow us to draw the following key conclusions. Here we write $z_x = \Phi^{-1}(x)$.

Theorem 1. *Suppose that $\delta = O(n^{-1/2})$ but $\delta \neq o(n^{-1/2})$. Then for the approximate p-values $\hat{P} = \hat{P}_{\mathrm{cB}}$ and $\hat{P} = \hat{P}_{\mathrm{B}}$,*

$$\begin{aligned}\mathrm{pr}_\eta\left(\hat{P} \leqslant x\right) &- \mathrm{pr}_\eta\left(\hat{P}_{\mathrm{N}} \leqslant x\right) \\ &= \left\{\mathrm{pr}_{\eta_0}\left(\hat{P}_{\mathrm{N}} \leqslant x\right) - x\right\}\left\{-\frac{\phi(z_x + n^{1/2}\check{\sigma}^{-1}|\theta - \theta_0|)}{\phi(z_x)} + O\left(n^{-1/2}\right)\right\}. \qquad (2)\end{aligned}$$

Theorem 2. *Suppose that $\delta = o(n^{-1/2})$. Then*

$$\begin{aligned}\mathrm{pr}_\eta&\left(\hat{P}_{\mathrm{cB}} \leqslant x\right) - \mathrm{pr}_\eta\left(\hat{P}_{\mathrm{N}} \leqslant x\right) \\ &= \left\{\mathrm{pr}_{\eta_0}\left(\hat{P}_{\mathrm{N}} \leqslant x\right) - x\right\}\left\{-1 + n^{1/2}\check{\sigma}^{-1}|\theta - \theta_0|z_x + O\left(n^{-1} + n\|\delta\|^2\right)\right\} \qquad (3)\end{aligned}$$

and

$$\begin{aligned}\mathrm{pr}_\eta&\left(\hat{P}_{\mathrm{B}} \leqslant x\right) - \mathrm{pr}_\eta\left(\hat{P}_{\mathrm{N}} \leqslant x\right) \\ &= \left\{\mathrm{pr}_{\eta_0}\left(\hat{P}_{\mathrm{N}} \leqslant x\right) - x\right\} \\ &\quad \times \left\{-1 + n^{1/2}\check{\sigma}^{-1}|\theta - \theta_0|z_x + \frac{\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{B}} \leqslant x) - x}{\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{N}} \leqslant x) - x} + O\left(n^{-1} + n\|\delta\|^2\right)\right\}, \qquad (4)\end{aligned}$$

where the ratio $\{\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{B}} \leqslant x) - x\}/\{\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{N}} \leqslant x) - x\}$ in (4) is $O(n^{-1/2})$.

For an approximate p-value $\hat{P}$, define $Q(\hat{P}, \alpha; \eta, \eta_0) = \mathrm{pr}_\eta(\hat{P} \leqslant \alpha) - \mathrm{pr}_{\eta_0}(\hat{P} \leqslant \alpha)$.

Theorem 3. *Suppose that $\delta = o(n^{-1/2})$. Then for the approximate p-values $\hat{P} = \hat{P}_{\mathrm{cB}}$ and $\hat{P} = \hat{P}_{\mathrm{B}}$,*

$$\begin{aligned}Q\left(\hat{P}, \alpha; \eta, \eta_0\right) &= Q\left(\hat{P}_{\mathrm{N}}, \alpha; \eta, \eta_0\right) + \left\{\mathrm{pr}_{\eta_0}\left(\hat{P}_{\mathrm{N}} \leqslant \alpha\right) - \alpha\right\} n^{1/2}\check{\sigma}^{-1}|\theta - \theta_0|z_\alpha \\ &\quad + O\left(n^{-(\beta+2)/2} + n^{-1/2}\|\delta\| + n\|\delta\|^2 + n^{(1-\beta)/2}\|\delta\|\right).\end{aligned}$$

In the asymptotic regime where $\delta = O(n^{-1/2})$, asymptotically the power functions of the constrained and unconstrained bootstraps, $\mathrm{pr}_\eta(\hat{P}_{\mathrm{cB}} \leqslant \alpha)$ and $\mathrm{pr}_\eta(\hat{P}_{\mathrm{B}} \leqslant \alpha)$, are equal.

In the asymptotic regime where $\delta = o(n^{-1/2})$, we deduce from expansions (3) and (4) that asymptotically the power functions of the two bootstrap tests differ. However, the changes in power, $Q(\hat{P}_{\mathrm{cB}}, \alpha; \eta, \eta_0)$ and $Q(\hat{P}_{\mathrm{B}}, \alpha; \eta, \eta_0)$, are the same. So, asymptotically speaking, the difference in power of the two methods is essentially defined by the difference in their sizes. More specifically, from (3) and (4) we see that the leading term in an asymptotic expansion of $\mathrm{pr}_\eta(\hat{P}_{\mathrm{cB}} \leqslant \alpha) - \mathrm{pr}_\eta(\hat{P}_{\mathrm{B}} \leqslant \alpha)$ is given by $\alpha - \mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{B}} \leqslant \alpha)$. For the typical situation where the test is based on a statistic distributed under H_0 as $N(0, 1)$ with an error of $O(n^{-1/2})$, this discrepancy between the powers of the two bootstrap procedures is of $O(n^{-1})$ under this asymptotic regime. If the unconstrained bootstrap yields a test which is conservative, so that its actual size is smaller than the nominal size α, then the constrained bootstrap test has an asymptotically higher power than the unconstrained bootstrap under such local alternatives; on the other hand, when the unconstrained bootstrap test is liberal, the asymptotic power of the constrained bootstrap test will be lower than for the unconstrained bootstrap. We demonstrate in § 4 that this asymptotic comparison predicts well the behaviours of the two bootstrap tests in terms of power as the hypothesized θ_0 moves away from the true value of θ in finite-sample contexts.

Further, while we have argued from Theorem 3 that the power functions of the two bootstrap tests grow at the same rate as θ_0 moves away from the true θ, we see that the power of the test based on normal approximation grows more slowly if its actual size is below the nominal size α, but more quickly if its actual size is above the nominal α. We are again able to provide in § 4 vivid illustration of this asymptotic behaviour for a small sample size n.

We can also consider bootstrap p-values based on simulation using estimates of the nuisance parameter other than the global and constrained maximum likelihood estimators; see, for instance, Severini (1998) and Yang et al. (2014). Let $\tilde{\eta}_\vartheta$ be a $n^{1/2}$-consistent estimator of η, constrained to satisfy the condition $g(\tilde{\eta}_\vartheta) = \vartheta$. Assume differentiability of the map $\vartheta \mapsto \tilde{\eta}_\vartheta$ around $\vartheta = \theta_0$ and that $\mathrm{cov}_\eta\{T(\theta_0), \tilde{\eta}_{\theta_0}\} = O(n^{-1/2})$ for $\delta = O(n^{-1/2})$, which holds under mild regularity conditions. Arguments analogous to those presented in the Supplementary Material can be used to deduce that (2) and (4) hold with $\hat{\eta}$ replaced by $\tilde{\eta}_{\theta_0}$. For the special case of $\tilde{\eta}_\vartheta = \hat{\eta}_\vartheta$, $\mathrm{cov}_\eta\{T(\theta_0), \hat{\eta}_{\theta_0}\}$ has a smaller order, $O(n^{-1})$, which leads to a different expansion (3) under $\delta = o(n^{-1/2})$.

Consider the particular case where $T(\theta_0) = R(\theta_0)$, the signed root likelihood ratio statistic, for which $\mathrm{pr}_{\eta_0}(\hat{P}_{\mathrm{N}} \leqslant x) = x + O(n^{-1/2})$. If $\delta = O(n^{-1/2})$ but $\delta \neq o(n^{-1/2})$, by Theorem 1 we have that $\mathrm{pr}_\eta(\hat{P}_{\mathrm{N}} \leqslant x)$, $\mathrm{pr}_\eta(\hat{P}_{\mathrm{cB}} \leqslant x)$ and $\mathrm{pr}_\eta(\hat{P}_{\mathrm{B}} \leqslant x)$ are asymptotically equivalent up to $O(n^{-1/2})$. For the case of $\delta = o(n^{-1/2})$, by Theorem 2 we have that

$$\mathrm{pr}_\eta\big(\hat{P}_{\mathrm{cB}} \leqslant x\big) - \mathrm{pr}_\eta\big(\hat{P}_{\mathrm{N}} \leqslant x\big) = \big\{x - \mathrm{pr}_{\eta_0}\big(\hat{P}_{\mathrm{N}} \leqslant x\big)\big\}\big\{1 + O\big(n^{-1} + n^{1/2}\|\delta\|\big)\big\}$$

and

$$\mathrm{pr}_\eta\big(\hat{P}_{\mathrm{B}} \leqslant x\big) - \mathrm{pr}_\eta\big(\hat{P}_{\mathrm{N}} \leqslant x\big) = \big\{x - \mathrm{pr}_{\eta_0}\big(\hat{P}_{\mathrm{N}} \leqslant x\big)\big\}\big\{1 + O\big(n^{-1/2} + n^{1/2}\|\delta\|\big)\big\}.$$

It is of interest to compare the two bootstrap approximations to $R(\theta_0)$ with normal approximation to the adjusted signed root statistic $R^*(\theta_0)$ (Barndorff-Nielsen, 1986). The p-value based on the analytic normal approximation to $R^*(\theta_0)$ is

$$\hat{P}_{\mathrm{A}} = 1 - \Phi\{R^*(\theta_0)\}.$$

Setting $\beta = 3$ in the technical derivations, we deduce that if $\delta = O(n^{-1/2})$ but $\delta \neq o(n^{-1/2})$,

$$\mathrm{pr}_\eta\big(\hat{P}_\mathrm{A} \leqslant x\big) = \mathrm{pr}_\eta\big(\hat{P}_\mathrm{N} \leqslant x\big) + O(n^{-1/2}),$$

and that if $\delta = o(n^{-1/2})$ we have

$$\mathrm{pr}_\eta\big(\hat{P}_\mathrm{A} \leqslant x\big) - \mathrm{pr}_\eta\big(\hat{P}_\mathrm{N} \leqslant x\big) = \big\{x - \mathrm{pr}_{\eta_0}\big(\hat{P}_\mathrm{N} \leqslant x\big)\big\}\big\{1 + O\big(n^{-1} + n^{1/2}\|\delta\| + n^{3/2}\|\delta\|^2\big)\big\}.$$

We see that the p-value based on normal approximation to $R^*(\theta_0)$ has a distribution which is asymptotically closer to that of the p-value based on constrained bootstrap approximation to $R(\theta_0)$ than the one based on the unconstrained bootstrap if the alternative is sufficiently local, with $\delta = o(n^{-1})$. For $\delta = o(n^{-1/2})$, the discrepancy between the power and size of the normal approximation to $R^*(\theta_0)$ has the expansion

$$\begin{aligned}\mathrm{pr}_\eta\big\{R^*(\theta_0) \geqslant z_{1-\alpha}\big\} &- \mathrm{pr}_{\eta_0}\big\{R^*(\theta_0) \geqslant z_{1-\alpha}\big\}\\ &= \mathrm{pr}_\eta\big(\hat{P}_\mathrm{N} \leqslant \alpha\big) - \mathrm{pr}_{\eta_0}\big(\hat{P}_\mathrm{N} \leqslant \alpha\big) + O\big(\|\delta\| + n\|\delta\|^2\big).\end{aligned}$$

The corresponding discrepancies for the two bootstrap approximations have the same expansion as above except for an additional $O(n^{-3/2})$ term, as can be deduced from Theorem 3.

4. EXAMPLES

4·1. *Inverse Gaussian mean*

Suppose that $Y = (Y_1, \ldots, Y_n)$ where $Y_1, \ldots, Y_n$ are independent, identically distributed inverse Gaussian random variables with mean θ and shape parameter λ, so that the common density is

$$f(y; \theta, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{-\frac{\lambda(y-\theta)^2}{2\theta^2 y}\right\} \quad (y > 0; \theta, \lambda > 0).$$

Inference is required for the mean θ, with λ a nuisance parameter. In this example, global and constrained maximum likelihood estimators have explicit, closed-form expressions, so no numerical optimization is needed in constructing the signed root statistic $R(\theta_0)$, which is used throughout our analysis, or its adjusted form $R^*(\theta_0)$. The power of tests based on normal approximation and the two bootstrap procedures are compared for nominal sizes $\alpha = 1\%$, 5% and 10% in Tables 1 and 2, for a range of sample sizes n. In all cases, the true parameter values are $\theta = \lambda = 2{\cdot}0$. Table 1 displays the results of testing $H_0 : \theta = \theta_0$ against $H_\mathrm{a} : \theta > \theta_0$, and Table 2 the results of testing against $H_\mathrm{a} : \theta < \theta_0$. All figures are based on 50 000 replications, with 20 000 samples being drawn in the calculation of each bootstrap p-value. The results are broadly as predicted by the theory. In particular, there is little discernible difference between the powers of the two bootstrap tests and the test based on normal approximation to the distribution of the adjusted signed root statistic, with the discrepancies reflecting slight differences in size for the small sample sizes n considered. Of particular interest, however, is the small-sample case $n = 5$ in Table 2. Here, the unconstrained bootstrap has actual size noticeably above the nominal size, and the constrained bootstrap is more accurate in terms of size, with the power functions reflecting this difference. The normal approximation to the distribution of $R(\theta_0)$ yields size substantially above the nominal level. Figure 1(a) shows a more complete picture of the power functions for the $\alpha = 5\%$ case

Table 1. *Comparison of p-values for the inverse Gaussian mean example with nominal sizes* $\alpha = 1\%, 5\%, 10\%$*;* $\hat{P}_{\rm N}$ *and* $\hat{P}_{\rm A}$ *are p-values obtained by normal approximation to the signed root statistic* $R(\theta_0)$ *and its adjusted form* $R^*(\theta_0)$*, respectively;* $\hat{P}_{\rm B}$ *and* $\hat{P}_{\rm cB}$ *are p-values obtained by unconstrained and constrained bootstrap approximation of the distribution of* $R(\theta_0)$*, respectively. All figures are based on* 50 000 *replications, with* 20 000 *samples being drawn in the calculation of each bootstrap p-value; the figures give percentages of the* 50 000 *p-values that are less than* α*, in testing against* $H_{\rm a} : \theta > \theta_0$

$\theta_0 =$		θ				$\theta - 2/n$				$\theta - 2/n^{1/2}$			
n	α	$\hat{P}_{\rm N}$	$\hat{P}_{\rm B}$	$\hat{P}_{\rm cB}$	$\hat{P}_{\rm A}$	$\hat{P}_{\rm N}$	$\hat{P}_{\rm B}$	$\hat{P}_{\rm cB}$	$\hat{P}_{\rm A}$	$\hat{P}_{\rm N}$	$\hat{P}_{B}$	$\hat{P}_{\rm cB}$	$\hat{P}_{\rm A}$
5	1%	1·5	0·8	0·9	1·0	5·5	3·4	3·7	4·0	25·8	19·6	20·5	21·5
	5%	5·5	5·0	4·9	5·2	15·5	14·3	14·1	14·6	48·4	45·9	45·7	46·7
	10%	9·7	10·2	9·8	10·2	23·7	24·1	23·5	24·1	59·8	59·5	59·1	59·7
10	1%	1·1	1·0	1·0	1·0	2·8	2·7	2·7	2·7	19·0	18·4	18·2	18·4
	5%	4·6	5·1	4·9	5·0	9·9	10·7	10·4	10·6	39·2	40·3	39·7	40·0
	10%	9·0	10·4	10·0	10·2	16·7	18·5	18·1	18·2	51·1	53·2	52·6	52·9
15	1%	1·0	1·1	1·0	1·1	2·2	2·3	2·3	2·3	16·8	17·2	16·9	17·1
	5%	4·5	5·1	4·9	5·0	8·3	9·3	9·0	9·1	36·1	37·9	37·4	37·5
	10%	8·9	10·2	10·0	10·0	14·9	16·7	16·3	16·4	48·2	50·4	50·0	50·2
20	1%	0·9	1·0	1·0	1·0	1·8	2·0	1·9	1·9	15·5	16·2	16·0	16·1
	5%	4·4	5·0	4·9	4·9	7·5	8·5	8·3	8·3	34·6	36·4	36·0	36·2
	10%	8·9	10·3	10·1	10·1	14·0	15·6	15·4	15·4	46·8	49·1	48·8	48·9
25	1%	0·9	1·0	0·9	0·9	1·7	1·9	1·8	1·9	14·5	15·3	15·1	15·2
	5%	4·5	5·1	5·0	5·0	7·3	8·1	8·0	8·0	33·0	34·9	34·5	34·6
	10%	9·0	10·2	10·0	10·1	13·2	14·7	14·5	14·6	45·6	47·8	47·5	47·6
50	1%	0·9	1·0	1·0	1·0	1·4	1·5	1·5	1·5	12·7	13·5	13·4	13·4
	5%	4·6	5·1	5·0	5·0	6·3	7·0	6·9	7·0	34·1	32·1	31·9	31·9
	10%	9·3	10·3	10·2	10·2	11·9	13·1	13·0	13·0	43·1	45·1	44·9	45·0

in this context: for each of the approximate p-values $\hat{P} = \hat{P}_{\rm B}$, $\hat{P}_{\rm cB}$ and $\hat{P}_{\rm A}$, the discrepancy

$$D = \{\mathrm{pr}_\eta(\hat{P} \leqslant \alpha) - \mathrm{pr}_\eta(\hat{P}_{\rm N} \leqslant \alpha)\} / |\alpha - \mathrm{pr}_{\eta_0}(\hat{P}_{\rm N} \leqslant \alpha)|,$$

which the theory of § 3 indicates is most relevant, is plotted against $\theta_0 - \theta$. The plot was constructed by interpolation of power values obtained from simulating, as before, at 11 values of θ_0, including those considered in Table 2. As the theory predicts, for each method the transformed power D lies close to a straight line with negative intercept and negative slope. The normal approximation is liberal, and the theory implies that the power for each method is both smaller and grows more slowly than that of the normal approximation. The unconstrained bootstrap is also liberal, and yields power greater than that of the constrained bootstrap, but which increases at the same rate, at least for very local departures from the null hypothesis. The power figures for the test based on $R^*(\theta_0)$ are closer to those of the constrained bootstrap than those of the unconstrained bootstrap, again giving empirical support to the theory presented above. Further numerical results are reported in the Supplementary Material.

A key observation is that in this model the signed root statistic is highly pivotal: since its distribution depends very little on the value of the nuisance parameter λ, there is relatively little practical difference between the distribution of p-values calculated by the unconstrained and constrained bootstraps. This is not necessarily the case for other likelihood-based statistics. Consider, for example, the Wald statistic, defined as $\hat{\theta} - \theta_0$, standardized by a variance estimate

Table 2. *Comparison of p-values for the inverse Gaussian mean example with* $n = 5, 10, 15$, *nominal sizes* $\alpha = 1\%, 5\%, 10\%$ *and* $\Delta = 1{\cdot}0$. *All figures are based on* 50 000 *replications, with* 20 000 *samples being drawn in the calculation of each bootstrap p-value; the figures give percentages of p-values that are less than* α, *in testing against* $H_{\rm a} : \theta < \theta_0$

$\theta_0 =$	θ	$\theta + \Delta$	$\theta + 2\Delta$	$\theta + 3\Delta$	$\theta + 4\Delta$
			$n = 5$		
$\hat{P}_{\rm N}$	(3·2, 11·0, 18·6)	(8·5, 26·1, 40·0)	(13·5, 38·1, 55·6)	(17·5, 47·0, 66·1)	(20·8, 53·4, 73·1)
$\hat{P}_{\rm B}$	(1·2, 5·8, 11·3)	(3·2, 14·1, 25·8)	(5·1, 21·4, 37·2)	(6·6, 27·1, 45·4)	(8·0, 31·3, 51·5)
$\hat{P}_{\rm cB}$	(1·0, 5·1, 10·0)	(2·5, 12·0, 22·9)	(3·9, 17·8, 32·6)	(5·0, 22·3, 39·5)	(5·9, 25·7, 44·5)
$\hat{P}_{\rm A}$	(1·0, 5·2, 10·4)	(2·7, 12·4, 23·4)	(4·2, 18·5, 33·2)	(5·4, 23·1, 40·3)	(6·3, 26·7, 45·4)
$\hat{P}_{\rm W}$	(18·2, 24·8, 29·6)	(43·8, 53·5, 60·0)	(63·8, 73·5, 78·9)	(77·2, 85·0, 89·0)	(85·7, 91·5, 94·2)
$\hat{P}_{\rm BW}$	(3·9, 9·9, 15·2)	(12·4, 26·0, 35·9)	(22·3, 40·7, 52·6)	(31·4, 52·3, 64·6)	(39·5, 61·4, 73·2)
$\hat{P}_{\rm cBW}$	(0·9, 5·1, 10·4)	(2·1, 12·0, 23·5)	(3·2, 17·7, 33·4)	(3·9, 22·1, 40·6)	(4·6, 25·4, 45·8)
			$n = 10$		
$\hat{P}_{\rm N}$	(2·1, 8·3, 14·8)	(10·5, 31·6, 48·0)	(21·2, 53·8, 72·5)	(31·3, 69·0, 85·5)	(39·4, 78·5, 91·9)
$\hat{P}_{\rm B}$	(1·1, 5·5, 10·8)	(6·0, 22·4, 37·3)	(12·8, 40·5, 60·1)	(19·3, 54·2, 74·5)	(24·9, 64·1, 82·9)
$\hat{P}_{\rm cB}$	(1·0, 5·1, 10·1)	(5·2, 20·6, 35·3)	(11·0, 37·1, 57·0)	(16·6, 50·0, 71·1)	(21·2, 58·8, 79·4)
$\hat{P}_{\rm A}$	(1·0, 5·2, 10·2)	(5·3, 20·7, 35·3)	(11·1, 37·2, 57·0)	(16·7, 50·0, 71·0)	(21·5, 58·9, 79·3)
$\hat{P}_{\rm W}$	(11·5, 18·2, 23·4)	(45·2, 58·0, 66·0)	(73·4, 83·5, 88·5)	(88·2, 94·1, 96·4)	(95·0, 98·0, 99·0)
$\hat{P}_{\rm BW}$	(2·8, 7·8, 12·8)	(15·1, 31·9, 44·5)	(32·7, 56·3, 70·0)	(49·3, 73·4, 84·3)	(62·5, 83·7, 91·7)
$\hat{P}_{\rm cBW}$	(1·0, 5·2, 10·3)	(4·9, 20·7, 35·8)	(10·1, 37·3, 57·7)	(15·1, 50·1, 71·7)	(19·2, 59·0, 80·0)
			$n = 15$		
$\hat{P}_{\rm N}$	(1·8, 7·6, 13·8)	(14·6, 39·6, 57·0)	(33·3, 69·2, 84·5)	(50·3, 85·1, 94·9)	(62·5, 92·7, 98·3)
$\hat{P}_{\rm B}$	(1·1, 5·5, 10·8)	(9·9, 31·6, 48·3)	(24·2, 58·9, 76·9)	(37·8, 76·4, 90·2)	(49·1, 86·0, 95·7)
$\hat{P}_{\rm cB}$	(1·1, 5·2, 10·4)	(9·0, 29·9, 46·7)	(21·9, 56·2, 75·1)	(34·5, 73·4, 88·8)	(44·8, 83·5, 94·6)
$\hat{P}_{\rm A}$	(1·1, 5·2, 10·4)	(9·1, 30·0, 46·7)	(22·0, 56·3, 74·9)	(34·5, 73·4, 88·7)	(44·9, 83·4, 94·5)
$\hat{P}_{\rm W}$	(9·3, 15·6, 20·8)	(49·6, 64·1, 71·9)	(81·7, 90·6, 94·2)	(94·8, 98·0, 99·0)	(98·6, 99·6, 99·9)
$\hat{P}_{\rm BW}$	(2·2, 7·1, 12·2)	(19·4, 39·5, 53·5)	(45·0, 70·5, 82·4)	(66·4, 87·1, 94·1)	(80·3, 94·6, 98·0)
$\hat{P}_{\rm cBW}$	(1·0, 5·3, 10·5)	(8·7, 30·1, 47·0)	(21·1, 56·5, 75·4)	(33·2, 73·6, 89·0)	(43·1, 83·6, 94·8)

calculated from the expected information evaluated at the global maximum likelihood estimator. Table 2 also reports results corresponding to the approximate p-values $\hat{P}_{\rm W}$, $\hat{P}_{\rm BW}$ and $\hat{P}_{\rm cBW}$ obtained, respectively, by normal approximation and the unconstrained and constrained bootstrap approaches applied with this Wald statistic. Now a substantial difference can be seen in the sampling distributions of the p-values $\hat{P}_{\rm BW}$ and $\hat{P}_{\rm cBW}$. Plots of the discrepancy D for this statistic are shown in the Supplementary Material.

In this problem, if the shape parameter λ is the parameter of interest, with the mean θ being the nuisance parameter, then both the global and the constrained maximum likelihood estimators of θ are $\bar{Y} = \sum_{i=1}^{n} Y_i/n$, and in fact the signed root statistic is exactly pivotal, so that the two bootstrap testing procedures coincide and yield tests of size exactly equal to the nominal desired size, modulo simulation error. However, this is not the case for the inference problem being considered here.

4·2. *Multisample normal model*

A more challenging example, considered by Sartori et al. (1999), involves a high-dimensional nuisance parameter. We observe Y_{ij} $(i = 1, \ldots, g;\ j = 1, \ldots, n)$, which are independent normal

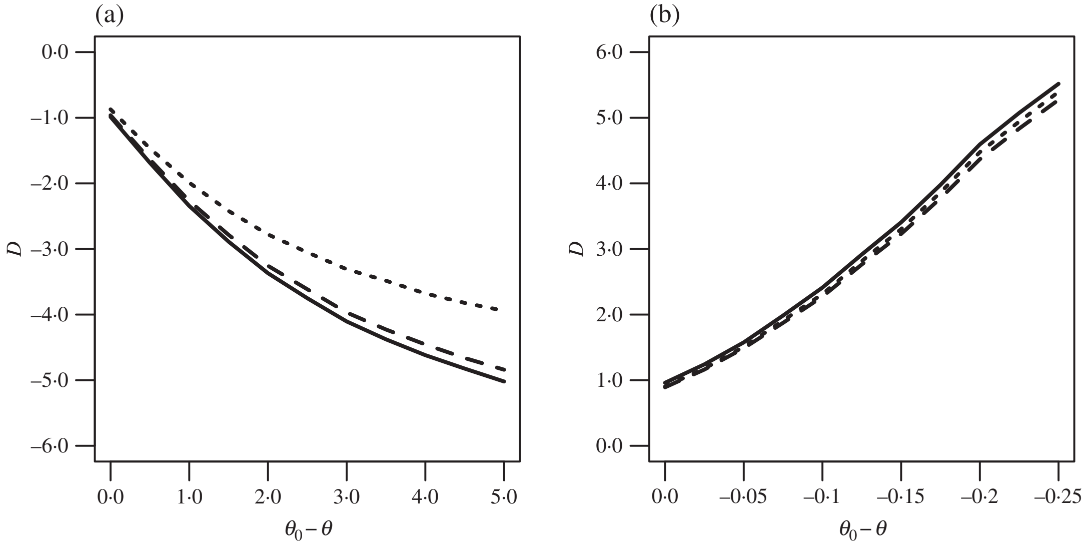

Fig. 1. The discrepancy D plotted against $\theta_0 - \theta$ in the case of nominal size $\alpha = 5\%$, for $\hat{P}_{\mathrm{cB}}$ (solid), $\hat{P}_{\mathrm{B}}$ (dotted) and $\hat{P}_{\mathrm{A}}$ (dashed): (a) inverse Gaussian example with $n = 5$, testing $H_0 : \theta = \theta_0$ against $H_{\mathrm{a}} : \theta < \theta_0$; (b) normal example with $n = 5$ and $g = 5$, testing $H_0 : \theta = \theta_0$ against $H_{\mathrm{a}} : \theta > \theta_0$.

Table 3. *Comparison of p-values for the normal example with nominal sizes* $\alpha = 1\%, 5\%, 10\%$ *and* $g = 5$. *The figures are based on* 50 000 *replications with* 20 000 *samples drawn in the calculation of each bootstrap p-value for* $n = 10, 20, 50$, *and based on* 10 000 *replications with* 10 000 *samples drawn for* $n = 100, 200, 500$; *the figures give percentages of p-values that are less than* α, *in testing against* $H_{\mathrm{a}} : \theta < \theta_0$

$\theta_0 =$		θ				$\theta + 0{\cdot}5/n$				$\theta + 0{\cdot}5/n^{1/2}$			
n	α	$\hat{P}_{\mathrm{N}}$	$\hat{P}_{\mathrm{B}}$	$\hat{P}_{\mathrm{cB}}$	$\hat{P}_{\mathrm{A}}$	$\hat{P}_{\mathrm{N}}$	$\hat{P}_{\mathrm{B}}$	$\hat{P}_{\mathrm{cB}}$	$\hat{P}_{\mathrm{A}}$	$\hat{P}_{\mathrm{N}}$	$\hat{P}_{\mathrm{B}}$	$\hat{P}_{\mathrm{cB}}$	$\hat{P}_{\mathrm{A}}$
10	1%	4·1	1·0	1·0	1·0	7·3	1·9	12·0	2·0	19·0	6·4	6·7	6·9
	5%	14·2	4·8	5·0	5·1	22·0	8·5	8·8	8·9	43·3	21·5	22·0	22·3
	10%	23·7	9·7	10·0	10·2	34·2	15·9	16·3	16·6	58·4	34·4	35·1	35·5
20	1%	2·6	1·0	1·0	1·0	4·3	1·6	1·7	1·7	16·0	7·7	7·9	8·0
	5%	10·7	4·8	4·9	4·9	15·3	7·6	7·7	7·7	38·8	24·0	24·3	24·4
	10%	18·6	9·9	10·0	10·1	25·4	14·5	14·6	14·7	53·5	37·4	37·6	37·8
50	1%	2·0	1·1	1·1	1·1	2·7	1·6	1·6	1·6	14·0	9·1	9·2	9·3
	5%	8·2	5·0	5·1	5·1	10·5	6·7	6·8	6·8	35·2	26·2	26·3	26·3
	10%	15·0	10·0	10·0	10·1	18·7	12·9	12·9	13·0	49·7	39·8	40·0	40·0
100	1%	1·5	0·9	0·9	0·9	2·0	1·2	1·2	1·2	12·6	9·1	9·2	9·1
	5%	6·8	4·8	4·8	4·8	8·3	6·0	6·0	6·0	33·0	27·2	27·2	27·1
	10%	12·8	9·5	9·5	9·5	15·4	11·5	11·5	11·5	47·1	40·1	40·1	40·2
200	1%	1·2	0·9	0·9	0·9	1·5	1·1	1·1	1·1	2·7	10·4	10·4	10·3
	5%	6·2	4·7	4·7	4·8	7·3	5·7	5·7	5·7	32·6	28·2	28·2	28·4
	10%	12·5	10·2	10·2	10·2	13·9	11·6	11·6	11·6	47·3	42·0	42·0	42·1
500	1%	1·0	0·8	0·8	0·8	1·3	1·0	1·0	1·0	11·7	10·2	10·2	10·2
	5%	5·6	4·8	4·8	4·8	6·3	5·3	5·3	5·3	30·9	28·1	28·1	28·2
	10%	11·2	9·7	9·7	9·7	12·0	10·5	10·6	10·6	45·0	41·8	41·8	41·9

random variables with means μ_i and variances $\theta\mu_i^{1/2}$. The parameter of interest is θ, with $(\mu_1, \ldots, \mu_g)$ being the nuisance parameter. We set $g = 5$ or 10 and $\mu_i = i$, with the true θ equal to 0·7. We consider testing $H_0 : \theta = \theta_0$ against the alternatives $H_{\mathrm{a}} : \theta < \theta_0$ and $H_{\mathrm{a}} : \theta > \theta_0$, again

Table 4. *Comparison of p-values for the normal example with* $n = 5, 10, 20$, *nominal sizes* $\alpha = 1\%, 5\%, 10\%$, $\Delta = 0{\cdot}05$ *and* $g = 5$. *The figures are based on* 50 000 *replications with* 20 000 *samples drawn in the calculation of each bootstrap p-value for* $n = 5, 10$, *and based on* 10 000 *replications with* 10 000 *samples drawn for* $n = 20$; *the figures give percentages of p-values that are less than* α, *in testing against* $H_a : \theta > \theta_0$

$\theta_0 =$	θ	$\theta - \Delta$	$\theta - 2\Delta$	$\theta - 3\Delta$	$\theta - 4\Delta$
$n = 5$					
$\hat{P}_N$	(0·2, 0·9, 2·2)	(0·3, 1·9, 4·1)	(0·9, 3·7, 7·4)	(1·9, 7·2, 12·6)	(4·3, 12·9, 20·6)
$\hat{P}_B$	(0·9, 4·6, 9·6)	(1·8, 8·0, 14·8)	(3·7, 13·1, 22·1)	(7·3, 20·7, 31·8)	(13·1, 31·1, 43·5)
$\hat{P}_{cB}$	(1·0, 4·8, 9·8)	(2·0, 8·3, 15·1)	(4·0, 13·5, 22·3)	(7·7, 21·1, 32·1)	(13·7, 31·5, 43·8)
$\hat{P}_A$	(0·9, 4·6, 9·4)	(1·9, 7·9, 14·4)	(3·7, 12·9, 21·6)	(7·4, 20·4, 31·2)	13·2, 30·6, 42·7)
$n = 10$					
$\hat{P}_N$	(0·3, 1·6, 3·6)	(0·9, 3·9, 7·8)	(2·4, 8·9, 15·7)	(6·4, 18·2, 28·1)	(15·0, 32·9, 45·0)
$\hat{P}_B$	(1·0, 4·8, 9·7)	(2·6, 10·1, 18·0)	(6·4, 19·3, 30·1)	(14·1, 32·9, 45·7)	(27·3, 50·5, 63·4)
$\hat{P}_{cB}$	(1·0, 4·9, 9·9)	(2·7, 10·3, 18·2)	(6·6, 19·5, 30·3)	(14·5, 33·2, 46·0)	(27·8, 50·9, 63·6)
$\hat{P}_A$	(1·0, 4·8, 9·7)	(2·7, 10·1, 17·9)	(6·5, 19·3, 30·0)	(14·3, 32·9, 45·6)	(27·6, 50·5, 63·2)
$n = 20$					
$\hat{P}_N$	(0·4, 2·3, 5·1)	(1·7, 7·5, 13·8)	(6·4, 19·4, 29·6)	(19·1, 39·2, 52·8)	(41·0, 65·1, 75·7)
$\hat{P}_B$	(1·0, 5·1, 10·3)	(3·8, 13·7, 22·9)	(12·1, 29·7, 42·5)	(28·5, 53·1, 66·0)	(54·0, 75·9, 84·4)
$\hat{P}_{cB}$	(1·1, 5·1, 10·4)	(3·8, 13·8, 23·0)	(12·2, 29·9, 42·6)	(28·8, 53·2, 66·1)	(54·2, 76·0, 84·5)
$\hat{P}_A$	(1·1, 5·1, 10·3)	(3·9, 13·8, 22·9)	(12·0, 29·7, 42·4)	(28·6, 53·0, 66·0)	(54.2, 75.8, 84.4)

using the signed root likelihood ratio statistic $R(\theta_0)$. Numerical results for tests of nominal size $\alpha = 1\%$, 5% and 10% are reported in Tables 3 and 4 for the case where $g = 5$; further results, which include the case of $g = 10$, are given in the Supplementary Material. Now the adjusted signed root statistic $R^*(\theta_0)$ is intractable and the analytic p-value $\hat{P}_A$ is based on the approximation described by Skovgaard (1996). Again, the results are as predicted by theory. Across all the scenarios studied, there is no substantial discrepancy in the power properties of the two bootstrap procedures, with the slight differences that are seen reflecting differences in actual size across the replications. There is close agreement between the results for the bootstrap p-values $\hat{P}_{cB}$ and those for the analytic p-values $\hat{P}_A$. Both bootstrap procedures are very accurate in terms of size, even in the context of a 10-dimensional nuisance parameter ($g = 10$). In this example, normal approximation to the unadjusted statistic delivers tests of actual size very different from the nominal desired size, being liberal for testing against $H_a : \theta < \theta_0$ and conservative for testing against $H_a : \theta > \theta_0$. Figure 1(b) shows a more complete picture of the power functions for the case in Table 4 where $n = 5$ and $g = 5$, plotting the discrepancy D against $\theta_0 - \theta$ for $\hat{P}_A$, $\hat{P}_B$ and $\hat{P}_{cB}$. The graphs were again obtained by interpolation from the simulated power values at 11 values of θ_0. Linearity and positive slope of the discrepancy D is again as predicted by the asymptotic theory. The normal approximation has size substantially below the nominal 5% considered here, and its power function increases more slowly as $\theta_0 - \theta$ decreases from 0 than do those of the bootstrap tests or the test based on normal approximation to the adjusted statistic, which are fairly indistinguishable in terms of power.

5. Discussion

Inference on a scalar parameter of interest in the presence of a nuisance parameter can conveniently be made using a likelihood-based test statistic which is asymptotically distributed as standard normal under a null hypothesis of interest. We have examined higher-order expansions of the distribution of p-values obtained by normal approximation and by bootstrap approximation under an asymptotic regime involving a general contiguous alternative hypothesis. Our analysis is based on the testing framework described by DiCiccio et al. (2001) and Lee & Young (2005). That framework, and the conclusions of Lee & Young (2005) concerning the distribution of p-values under the null hypothesis, was extended by Stern (2006) to test statistics based on a certain class of M-estimators, and future extension of the results here concerning distributions of p-values under an alternative hypothesis to such statistics would be worthwhile.

In the literature there is relatively little on finite-sample comparisons of the distributions of different approximate p-values under an alternative hypothesis, although some evidence has been provided for very specific cases; see, for example, Hung et al. (1997). Martin (2007) provided empirical results on the power of bootstrap tests when applied to common statistical inference problems. A general first-order asymptotic analysis of the sampling distributions of various p-values, primarily those motivated by Bayesian considerations, is given by Robins et al. (2000). Among the methods they consider is the constrained bootstrap p-value $\hat{P}_{\mathrm{cB}}$. The conditions on the test statistic assumed in our analysis ensure, in the language of Robins et al. (2000), an asymptotic frequentist p-value. It is readily established that the quantities of asymptotic relative power and asymptotic relative efficiency used by Robins et al. (2000) to distinguish between different p-value constructions coincide for all the p-values $\hat{P}_{\mathrm{N}}$, $\hat{P}_{\mathrm{cB}}$, $\hat{P}_{\mathrm{B}}$ and $\hat{P}_{\mathrm{A}}$ considered here, and higher-order analysis of p-values is therefore necessary to provide asymptotic discrimination between the different approximate p-values. Of particular interest is the elucidation of the asymptotic behaviour under an alternative hypothesis of constrained and unconstrained bootstrap p-values, $\hat{P}_{\mathrm{cB}}$ and $\hat{P}_{\mathrm{B}}$. Importantly from a methodological perspective, the asymptotic analysis is found to predict well the distribution of p-values observed for small sample sizes n. The comparative power properties of the two bootstrap procedures are seen to reflect the respective discrepancies between the actual sizes of the tests and the nominal desired size, which are often quite negligible in practice (Lee & Young, 2005; Young & Smith, 2005, Ch. 11).

Acknowledgement

This research was supported in part by the European Cooperation in Science and Technology.

Supplementary material

Supplementary material available at *Biometrika* online includes derivations of the technical results described in § 3 and further numerical results for both of the examples in § 4.

References

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

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.

DiCiccio, T. J., Kuffner, T. A., Young, G. A. & Zaretzki, R. (2015). Stability and uniqueness of p-values for likelihood-based inference. *Statist. Sinica* **25**, 1355–76.

Hung, H. M. J., O'Neill, R. T., Bauer, P. & Köhne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics* **53**, 11–22.

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

Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Comp. Statist. Data Anal.* **51**, 6321–42.

Robins, J. M., van der Vaart, A. & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *J. Am. Statist. Assoc.* **95**, 1143–56.

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

Severini, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507–22.

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

Stern, S. E. (2006). Simple and accurate one-sided inference based on a class of M-estimators. *Biometrika* **93**, 973–87.

Yang, N., Lian, K.-Y. & Reid, N. (2014). Reducing the sensitivity to nuisance parameters in pseudo-likelihood functions. *Can. J. Statist.* **42**, 544–62.

Young, G. A. & Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge: Cambridge University Press.

[*Received September* 2015. *Revised May* 2016]

Biometrika, *pp.* 1–10
© 2016 Biometrika Trust
Printed in Great Britain

Supplementary material for Distribution of likelihood-based p-values under a local alternative hypothesis

BY STEPHEN M.S. LEE

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
smslee@hku.hk

AND G. ALASTAIR YOUNG

Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.
alastair.young@imperial.ac.uk

1. TECHNICAL DETAILS

Derivations are given of the theoretical results presented in §3 of the main text.

Define $H^b(\eta) = \sigma^{-1}(\eta)g_a(\eta)J^{ab}(\eta,\eta)$, $H^b_i(\eta) = \partial H^b(\eta)/\partial\eta^i$, $H^b_{ij}(\eta) = \partial^2 H^b(\eta)/\partial\eta^i\partial\eta^j$, etc. It follows by Taylor expanding about η that

$$T(\theta_0) = T(\theta) \mp (M_n - \mu_n) \mp \mu_n + O_p\left(n^{1/2}\|\delta\|^3\right), \tag{S1}$$

where

$$\begin{aligned} M_n = {} & n^{1/2}\left\{H^b(\eta)s_{bi}(\eta) + H^b_i(\eta)s_b(\eta)\right\}\delta^i \\ & - n^{1/2}(1/2)\left\{H^b_{ij}(\eta)s_b(\eta) + H^b(\eta)s_{bij}(\eta) + 2H^b_i(\eta)s_{bj}(\eta)\right\}\delta^i\delta^j \end{aligned}$$

and

$$\mu_n = E_\eta M_n = -\,n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i - n^{1/2}(1/2)\left\{H^b(\eta)L_{bij}(\eta) - 2H^b_i(\eta)J_{bj}(\eta,\eta)\right\}\delta^i\delta^j.$$

Denote by $G(\cdot;\eta,\theta_0)$ the distribution function of $T(\theta_0)$. As in Lee & Young (2005), we assume that $G(\cdot;\eta,\theta)$ has the expansion

$$G(x;\eta,\theta) = \Phi(x) + n^{-\beta/2}d_n(\eta,x)\phi(x), \tag{S2}$$

for some $\beta \in \{1,2,\ldots\}$ and $O(1)$ function $d_n(\cdot)$, with Φ and ϕ denoting the standard normal distribution and density functions respectively. Inverting (S2) gives

$$G^{-1}(x;\eta,\theta) = z_x - n^{-\beta/2}d_n(\eta,z_x) - n^{-\beta}d_n(\eta,z_x)\left\{z_x d_n(\eta,z_x)/2 - d'_n(\eta,z_x)\right\} + O(n^{-3\beta/2}), \tag{S3}$$

where $z_x = \Phi^{-1}(x)$ and $d'_n(\eta,x) = \partial d_n(\eta,x)/\partial x$. Define $L_{b,ij}(\eta) = n\,E_\eta\left\{s_b(\eta)s_{ij}(\eta)\right\}$, which has order $O(1)$, and $K_i(\eta) = \sigma(\eta)^{-1}g_a(\eta)\left\{J^{aj}(\eta,\eta)L_{j,bi}(\eta)H^b(\eta) + H^a_i(\eta)\right\}$. Then we have, by noting that $\mathrm{cov}_\eta\left\{T(\theta), s_i(\eta)\right\} = \pm\,n^{-1/2}\sigma(\eta)^{-1}g_i(\eta) + O(n^{-1})$ and

$$\mathrm{cov}_\eta\left\{T(\theta), s_{ij}(\eta)\right\} = \pm\,n^{-1/2}\sigma(\eta)^{-1}g_a(\eta)J^{ab}(\eta,\eta)L_{b,ij}(\eta) + O(n^{-1}),$$

that

$$\mathrm{var}_\eta\left\{T(\theta) \mp (M_n - \mu_n)\right\} - \mathrm{var}_\eta\left\{T(\theta)\right\} = -\,2K_i(\eta)\delta^i + O\left(n^{-1/2}\|\delta\|\right).$$

Similar arguments show that the third- and higher-order cumulants of $T(\theta)$ and $T(\theta) \mp (M_n - \mu_n)$ differ by order $O\left(n^{-1/2}\|\delta\|\right)$. It then follows by (S1) and the delta method that

$$\begin{aligned} G(x;\eta,\theta_0) &= \mathrm{pr}_\eta\left\{T(\theta) \mp (M_n - \mu_n) \le x \pm \mu_n\right\} + O\left(n^{1/2}\|\delta\|^3\right) \\ &= G(x \pm \mu_n;\eta,\theta) + K_i(\eta)\delta^i(x \pm \mu_n)\phi(x \pm \mu_n) + O\left(n^{-1/2}\|\delta\|\right). \qquad \text{(S4)} \end{aligned}$$

P-value based on normal approximation

As noted before, the asymptotic $N(0,1)$ distribution of $T(\theta_0)$ under the null hypothesis allows the p-value for the test to be approximated by $\widehat{P}_{\mathrm{N}} = 1 - \Phi\{T(\theta_0)\}$. It follows by (S2) and (S4) that the distribution function of $\widehat{P}_{\mathrm{N}}$ has the expansion

$$\begin{aligned} \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) &= 1 - G(-z_x;\eta,\theta_0) \\ &= \Phi(z_x \mp \mu_n) - n^{-\beta/2}d_n(\eta, -z_x \pm \mu_n)\phi(z_x \mp \mu_n) \\ &\quad + K_i(\eta)\delta^i(z_x \mp \mu_n)\phi(z_x \mp \mu_n) + O\left(n^{-1/2}\|\delta\|\right). \qquad \text{(S5)} \end{aligned}$$

We consider two cases.

(i) If $\delta = O(n^{-1/2})$ but $\delta \ne o(n^{-1/2})$, the expansion (S5) reduces to

$$\begin{aligned} &\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) = \Phi\left\{z_x \pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i\right\} + \phi\left\{z_x \pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i\right\} \\ &\quad \times \Big[\pm n^{1/2}(1/2)\left\{H^b(\eta)L_{bij}(\eta) - 2H^b_i(\eta)J_{bj}(\eta,\eta)\right\}\delta^i\delta^j - n^{-\beta/2}d_n(\eta, -z_x \mp n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i) \\ &\qquad + K_a(\eta)\delta^a(z_x \pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i)\Big] + O(n^{-1}). \end{aligned}$$

(ii) If $\delta = o(n^{-1/2})$, the expansion (S5) reduces to

$$\begin{aligned} \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) &= x - n^{-\beta/2}d_n(\eta, -z_x)\phi(z_x) \pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i\phi(z_x) \\ &\quad + K_i(\eta)\delta^i z_x\phi(z_x) + O\left(n^{-1/2}\|\delta\| + n\|\delta\|^2 + n^{(1-\beta)/2}\|\delta\|\right). \qquad \text{(S6)} \end{aligned}$$

P-value based on constrained parametric bootstrap

Recall that the constrained parametric bootstrap estimates the null distribution of $T(\theta_0)$ by the bootstrap distribution of $T(Y^*_{\theta_0}, \theta_0)$, where $Y^*_{\theta_0}$ denotes a random sample of n observations drawn from $F_{\widehat{\eta}_{\theta_0}}$. Using (S2) and expanding about η, the above bootstrap distribution has the expansion

$$G(x;\widehat{\eta}_{\theta_0},\theta_0) = G(x;\eta,\theta) + n^{-\beta/2}\phi(x)\, d_{n,i}(\eta,x)(\widehat{\eta}_{\theta_0} - \eta)^i + O_p\left(n^{-\beta/2}\|\widehat{\eta}_{\theta_0} - \eta\|^2\right), \qquad \text{(S7)}$$

where $d_{n,i}(\eta,x) = \partial d_n(\eta,x)/\partial\eta^i$ for $i = 1,\ldots,d$. It can be shown, using the Lagrangian method and the fact $E_\eta(\check{s}_i) = \check{J}_{ij}\delta^j + O(\|\delta\|^2)$, that

$$\widehat{\eta}^i_{\theta_0} - \eta^i = \left(\check{J}^{ij} - \check{\sigma}^{-2}\check{g}_a\check{g}_b\check{J}^{ai}\check{J}^{bj}\right)\left\{\check{s}_j - E_\eta(\check{s}_j)\right\} - \check{\sigma}^{-2}\check{g}_a\check{J}^{ij}\check{g}_j\delta^a + O_p\left(n^{-1}\right) = O_p\left(n^{-1/2}\right). \qquad \text{(S8)}$$

Inverting (S7) and using (S8), we get

$$\begin{aligned} G^{-1}(x;\widehat{\eta}_{\theta_0},\theta_0) &= G^{-1}(x;\eta,\theta) - n^{-\beta/2}d_{n,i}(\eta,z_x)\left(\check{J}^{ij} - \check{\sigma}^{-2}\check{g}_a\check{g}_b\check{J}^{ai}\check{J}^{bj}\right)\left\{\check{s}_j - E_\eta(\check{s}_j)\right\} \\ &\quad + n^{-\beta/2}\check{\sigma}^{-2}d_{n,i}(\eta,z_x)\check{J}^{ij}\check{g}_j\check{g}_a\delta^a + O_p\left(n^{-(\beta+2)/2}\right). \qquad \text{(S9)} \end{aligned}$$

As defined previously, the constrained bootstrap estimate of the p-value is given by

$$\widehat{P}_{\mathrm{cB}} = 1 - G\{T(\theta_0);\widehat{\eta}_{\theta_0},\theta_0\}.$$

It follows by (S9) and the delta method that the sampling distribution of $\widehat{P}_{\mathrm{cB}}$ has the expansion

$$\begin{aligned}\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{cB}} \le x\right) &= \mathrm{pr}_\eta\left\{T(\theta_0) \ge G^{-1}(1-x;\widehat{\eta}_{\theta_0},\theta_0)\right\} \\ &= \mathrm{pr}_\eta\left\{T(\theta_0) + n^{-\beta/2}\widehat{A}_n(x) \ge G^{-1}(1-x;\eta,\theta) + n^{-\beta/2}\check{\sigma}^{-2}d_{n,i}(\eta,-z_x)\check{J}^{ij}\check{g}_j\check{g}_a\delta^a\right\} \\ &\qquad + O\left(n^{-(\beta+2)/2}\right), \qquad \text{(S10)}\end{aligned}$$

where $\widehat{A}_n(x) = d_{n,i}(\eta,-z_x)\big(\check{J}^{ij} - \check{\sigma}^{-2}\check{g}_a\check{g}_b\check{J}^{ai}\check{J}^{bj}\big)\left\{\check{s}_j - E_\eta(\check{s}_j)\right\}$. Noting that $\check{\sigma}^{-2}\check{g}_b\check{g}_j\check{J}^{bj} = 1$ and that $\mathrm{cov}_\eta\left\{T(\theta_0),\check{s}_j\right\} = \pm n^{-1/2}\check{\sigma}^{-1}\check{g}_j + O(n^{-1})$, we have

$$\begin{aligned}&\mathrm{var}_\eta\left\{T(\theta_0) + n^{-\beta/2}\widehat{A}_n(x)\right\} - \mathrm{var}_\eta\{T(\theta_0)\} \\ &\quad = \pm 2n^{-(\beta+1)/2}d_{n,i}(\eta,-z_x)\big(\check{J}^{ij} - \check{\sigma}^{-2}\check{g}_a\check{g}_b\check{J}^{ai}\check{J}^{bj}\big)\check{\sigma}^{-1}\check{g}_j + O(n^{-(\beta+2)/2}) = O(n^{-(\beta+2)/2}).\end{aligned}$$

Similarly, it can be shown that third- and higher-order cumulants of $T(\theta_0) + n^{-\beta/2}\widehat{A}_n(x)$ and $T(\theta_0)$ differ by $O(n^{-(\beta+2)/2})$, so that their respective distributions differ by the same order. Write $\tilde{z} = z_x \mp \mu_n$. It follows by (S2), (S3), (S4), (S5) and (S10) that

$$\begin{aligned}&\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{cB}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) \\ &= G(-z_x;\eta,\theta_0) - G\left\{G^{-1}(1-x;\eta,\theta) + n^{-\beta/2}\check{\sigma}^{-2}d_{n,i}(\eta,-z_x)\check{J}^{ij}\check{g}_j\check{g}_a\delta^a;\eta,\theta_0\right\} + O\left(n^{-(\beta+2)/2}\right) \\ &= -\left\{G'(-\tilde{z};\eta,\theta) + K_i(\eta)\delta^i(1-\tilde{z}^2)\phi(\tilde{z})\right\}\left\{z_x + G^{-1}(1-x;\eta,\theta)\right\} \\ &\quad - n^{-\beta/2}\check{\sigma}^{-2}d_{n,i}(\eta,-z_x)\check{J}^{ij}\check{g}_j\check{g}_a\delta^a\phi(\tilde{z}) - (1/2)\,\tilde{z}\phi(\tilde{z})\left\{z_x + G^{-1}(1-x;\eta,\theta)\right\}^2 + O\left(n^{-(\beta+2)/2}\right) \\ &= n^{-\beta/2}d_n(\eta,-z_x)\phi(\tilde{z}) - n^{-\beta}d_n(\eta,-z_x)\phi(\tilde{z}) \\ &\qquad\qquad \times\left\{(\tilde{z}+z_x)d_n(\eta,-z_x)/2 - \tilde{z}d_n(\eta,-\tilde{z}) + d_n'(\eta,-z_x) - d_n'(\eta,-\tilde{z})\right\} \\ &\quad - n^{-\beta/2}\left\{\sigma(\eta)^{-2}d_{n,i}(\eta,-z_x)J^{ij}(\eta,\eta)g_j(\eta)g_a(\eta) - d_n(\eta,-z_x)(1-\tilde{z}^2)K_a(\eta)\right\}\delta^a\phi(\tilde{z}) \\ &\quad + O\left(n^{-(\beta+2)/2}\right). \qquad \text{(S11)}\end{aligned}$$

The expansion (S11) can be simplified under two separate conditions on δ.

(i) If $\delta = O(n^{-1/2})$ but $\delta \ne o(n^{-1/2})$, we have

$$\begin{aligned}&\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{cB}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) \\ &= n^{-\beta/2}d_n(\eta,-z_x)\,\phi\left\{z_x \pm n^{1/2}\sigma(\eta)^{-1}g_c(\eta)\delta^c\right\} + O\left(n^{-(\beta+1)/2}\right).\end{aligned}$$

In general, the test statistic $T(\theta_0)$ is chosen such that the ± sign above corresponds to testing against the alternative $\theta > \theta_0$ and $\theta < \theta_0$ respectively. Noting that $\theta - \theta_0 = g_i(\eta_0)\delta^i + O(\|\delta\|^2)$ and applying (S6) under $\eta = \eta_0$, the expansion (2) follows.

(ii) If $\delta = o(n^{-1/2})$, the expansion (S11) reduces to

$$\begin{aligned}&\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{cB}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right) \\ &= n^{-\beta/2}\left\{1 \mp n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i z_x\right\}d_n(\eta,-z_x)\phi(z_x) + O\left(n^{-(\beta+2)/2} + n^{1-\beta/2}\|\delta\|^2\right),\end{aligned}$$

which implies (3).

P-value based on unconstrained parametric bootstrap

As before, denote by Y^* a random sample of n observations drawn from $F_{\widehat{\eta}}$. Recall that the unconstrained parametric bootstrap estimates the null distribution of $T(\theta_0)$ by the bootstrap distribution of

$T(Y^*,\widehat{\theta})$, which has the expansion

$$G(x;\widehat{\eta},\widehat{\theta}) = G(x;\eta,\theta) + n^{-\beta/2}\phi(x)\, d_{n,i}(\eta,x)(\widehat{\eta}-\eta)^i + O_p\left(n^{-\beta/2}\|\widehat{\eta}-\eta\|^2\right). \tag{S12}$$

Lagrangian arguments show that

$$\widehat{\eta}^i - \eta^i = \check{J}^{ij}\left\{\check{s}_j - E_\eta(\check{s}_j)\right\} + O_p\left(n^{-1}\right) = O_p\left(n^{-1/2}\right). \tag{S13}$$

Inverting (S12) and using (S13), we get

$$G^{-1}(x;\widehat{\eta},\widehat{\theta}) = G^{-1}(x;\eta,\theta) - n^{-\beta/2} d_{n,i}(\eta,z_x)\check{J}^{ij}\left\{\check{s}_j - E_\eta(\check{s}_j)\right\} + O_p\left(n^{-(\beta+2)/2}\right). \tag{S14}$$

It follows by (S14) and the delta method that

$$\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{B}} \le x\right) = \mathrm{pr}_\eta\left\{T(\theta_0) + n^{-\beta/2}\widehat{B}_n(x) \ge G^{-1}(1-x;\eta,\theta)\right\} + O\left(n^{-(\beta+2)/2}\right), \tag{S15}$$

where $\widehat{B}_n(x) = d_{n,i}(\eta,-z_x)\check{J}^{ij}\left\{\check{s}_j - E_\eta(\check{s}_j)\right\}$. Note that

$$\mathrm{var}\left\{T(\theta_0) + n^{-\beta/2}\widehat{B}_n(x)\right\} - \mathrm{var}\{T(\theta_0\} = \pm\, 2n^{-(\beta+1)/2}\check{\sigma}^{-1} d_{n,i}(\eta,-z_x)\check{g}_j\check{J}^{ij} + O(n^{-(\beta+2)/2}). \tag{S16}$$

On the other hand, the third- and higher-order cumulants of $T(\theta_0) + n^{-\beta/2}\widehat{B}_n(x)$ and $T(\theta_0)$ differ by $O(n^{-(\beta+2)/2})$. It then follows by (S3), (S4), (S15) and (S16) that

$$\begin{aligned}\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{B}} \le x\right) &= 1 - G\left\{G^{-1}(1-x;\eta,\theta)\,;\eta,\theta_0\right\}\\ &\quad \mp n^{-(\beta+1)/2}\check{\sigma}^{-1} d_{n,i}(\eta,-z_x)\check{g}_j\check{J}^{ij} z_x\phi(z_x) + O\left(n^{-(\beta+2)/2}\right).\end{aligned} \tag{S17}$$

Analogous to (S11), we have by subtracting (S5) from (S17) that

$$\begin{aligned}&\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{B}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right)\\ &= G(-z_x;\eta,\theta_0) - G\left\{G^{-1}(1-x;\eta,\theta)\,;\eta,\theta_0\right\}\\ &\quad \mp n^{-(\beta+1)/2}\check{\sigma}^{-1} d_{n,i}(\eta,-z_x)\check{g}_j\check{J}^{ij} z_x\phi(z_x) + O\left(n^{-(\beta+2)/2}\right)\\ &= -\left\{G'(-\tilde{z};\eta,\theta) + K_i(\eta)\delta^i(1-\tilde{z}^2)\phi(\tilde{z})\right\}\left\{z_x + G^{-1}(1-x;\eta,\theta)\right\}\\ &\quad - (1/2)\,\tilde{z}\phi(\tilde{z})\left\{z_x + G^{-1}(1-x;\eta,\theta)\right\}^2 \mp n^{-(\beta+1)/2}\check{\sigma}^{-1} d_{n,i}(\eta,-z_x)\check{g}_j\check{J}^{ij} z_x\phi(z_x)\\ &\qquad + O\left(n^{-(\beta+2)/2}\right)\\ &= n^{-\beta/2} d_n(\eta,-z_x)\phi(\tilde{z}) - n^{-\beta} d_n(\eta,-z_x)\phi(\tilde{z})\\ &\qquad\qquad \times\left\{(\tilde{z}+z_x)d_n(\eta,-z_x)/2 - \tilde{z}d_n(\eta,-\tilde{z}) + d_n'(\eta,-z_x) - d_n'(\eta,-\tilde{z})\right\}\\ &\quad + n^{-\beta/2} d_n(\eta,-z_x)(1-\tilde{z}^2)K_i(\eta)\delta^i\phi(\tilde{z})\\ &\quad \mp n^{-(\beta+1)/2}\sigma(\eta)^{-1} d_{n,i}(\eta,-z_x)g_j(\eta)J^{ij}(\eta,\eta)z_x\phi(z_x) + O\left(n^{-(\beta+2)/2}\right).\end{aligned} \tag{S18}$$

Thus the following holds for the unconstrained bootstrap.

(i) If $\delta = O(n^{-1/2})$ but $\delta \ne o(n^{-1/2})$, the expansion (S18) reduces to

$$\begin{aligned}&\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{B}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right)\\ &= n^{-\beta/2} d_n(\eta,-z_x)\,\phi\left\{z_x \pm n^{1/2}\sigma(\eta)^{-1}g_c(\eta)\delta^c\right\} + O\left(n^{-(\beta+1)/2}\right),\end{aligned}$$

which implies (2) of the main text.

(ii) If $\delta = o(n^{-1/2})$, the expansion (S18) reduces to

$$\begin{aligned}\mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{B}} \le x\right) - \mathrm{pr}_\eta\left(\widehat{P}_{\mathrm{N}} \le x\right)&\\ = n^{-\beta/2}\left\{1 \mp n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i z_x\right\} d_n(\eta, -z_x)\phi(z_x)&\\ \mp n^{-(\beta+1)/2}\sigma(\eta)^{-1}d_{n,i}(\eta,-z_x)g_j(\eta)J^{ij}(\eta,\eta)z_x\phi(z_x)&\\ + O\left(n^{-(\beta+2)/2} + n^{1-\beta/2}\|\delta\|^2\right).&\end{aligned} \tag{S19}$$

Then (4) follows by recalling (3) and applying (S6) and (S19) under $\eta = \eta_0$.

Change in power function

It may be of interest to compare the power under a local alternative with the actual size of a nominal level α test constructed by each of the three methods. In what follows we consider only a local alternative with $\delta = o(n^{-1/2})$.

(a) For the normal approximation method, we have

$$\begin{aligned}Q(\widehat{P}_{\mathrm{N}}, \alpha; \eta, \eta_0) = &\pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i\phi(z_\alpha) + K_i(\eta)\delta^i(z_\alpha)\phi(z_\alpha)\\ &+ O\left(n^{-1/2}\|\delta\| + n\|\delta\|^2 + n^{(1-\beta)/2}\|\delta\|\right).\end{aligned}$$

(b) For the constrained and unconstrained bootstrap methods, we have $Q(\widehat{P}_{\mathrm{cB}}, \alpha; \eta, \eta_0)$ and $Q(\widehat{P}_{\mathrm{B}}, \alpha; \eta, \eta_0)$ both equal to

$$\begin{aligned}&\pm n^{1/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i\phi(z_\alpha) \mp n^{(1-\beta)/2}\sigma^{-1}(\eta)g_i(\eta)\delta^i z_\alpha d_n(\eta, -z_\alpha)\phi(z_\alpha)\\ &+ K_i(\eta)\delta^i(z_\alpha)\phi(z_\alpha) + O\left(n^{-(\beta+2)/2} + n^{-1/2}\|\delta\| + n\|\delta\|^2 + n^{(1-\beta)/2}\|\delta\|\right),\end{aligned}$$

which proves Theorem 3.

2. Additional numerical results

Further Tables are presented for Examples 1 and 2 of the main text, and a Figure is provided of the discrepancy quantity D for the case of Fig. 1(a) of the main text, but when the statistic is the Wald statistic.

Table S1. *Comparison of* $p-$*values, inverse Gaussian mean example,* $n = 5, 10, 15$*, nominal sizes* $\alpha = (1, 5, 10)\%$*,* $\Delta = 0.25$*. All figures based on 50,000 replications, with 20,000 samples being drawn in calculation of each bootstrap* $p-$*value. Figures give percentages of p-values* $< \alpha$*. Testing against* $H_a : \theta > \theta_0$*.*

$\theta_0 =$	θ	$\theta - \Delta$	$\theta - 2\Delta$	$\theta - 3\Delta$	$\theta - 4\Delta$	$\theta - 5\Delta$
			$n = 5$			
$\widehat{P}_{\mathrm{N}}$	(1.5, 5.5, 9.7)	(3.3, 10.6, 17.1)	(7.5, 19.8, 29.3)	(16.8, 35.8, 47.1)	(35.0, 58.7, 69.5)	(65.0, 84.3, 90.2)
$\widehat{P}_{\mathrm{B}}$	(0.8, 5.0, 10.2)	(2.0, 9.6, 17.6)	(5.0, 18.5, 29.7)	(11.9, 33.7, 47.2)	(27.5, 56.3, 69.1)	(56.3, 82.3, 89.6)
$\widehat{P}_{\mathrm{cB}}$	(0.9, 4.9, 9.8)	(2.2, 9.5, 17.1)	(5.4, 18.3, 29.0)	(12.6, 33.5, 46.6)	(28.8, 56.2, 68.8)	(57.8, 82.3, 89.5)
$\widehat{P}_{\mathrm{A}}$	(1.0, 5.2, 10.2)	(2.4, 9.9, 17.6)	(5.8, 18.9, 29.8)	(13.3, 34.4, 47.30	(30.1, 57.0, 69.4)	59.2, 82.9, 89.8)
			$n = 10$			
$\widehat{P}_{\mathrm{N}}$	(1.1, 4.6, 9.0)	(3.5, 11.8, 19.4)	(10.9, 26.9, 38.2)	(29.4, 52.0, 63.9)	(61.4, 80.8, 87.7)	(91.3, 97.4, 98.7)
$\widehat{P}_{\mathrm{B}}$	(1.0, 5.1, 10.4)	(3.4, 12.6, 21.3)	(10.6, 28.1, 40.3)	(28.5, 53.0, 65.4)	(60.2, 81.1, 88.3)	(91.0, 97.4, 98.6)
$\widehat{P}_{\mathrm{cB}}$	(1.0, 4.9, 10.0)	(3.3, 12.3, 20.9)	(10.4, 27.6, 39.8)	(28.4, 52.5, 65.0)	(60.0, 80.9, 88.1)	(91.0, 97.4, 98.8)
$\widehat{P}_{\mathrm{A}}$	(1.0, 5.0, 10.2)	(3.4, 12.4, 21.1)	(10.6, 27.8, 40.0)	(28.6, 52.8, 65.2)	(60.4, 81.1, 88.2)	91.1, 97.4, 98.8)
			$n = 15$			
$\widehat{P}_{\mathrm{N}}$	(1.0, 4.5, 8.9)	(4.2, 13.8, 22.4)	(15.6, 34.2, 46.4)	(42.6, 65.5, 75.7)	(79.5, 91.7, 95.3)	(98.3, 99.6, 99.8)
$\widehat{P}_{\mathrm{B}}$	(1.1, 5.1, 10.2)	(4.4, 15.0, 24.6)	(16.0, 36.0, 48.6)	(42.9, 66.8, 77.2)	(79.5, 92.0, 95.6)	(98.3, 99.6, 99.9)
$\widehat{P}_{\mathrm{cB}}$	(1.0, 4.9, 10.0)	(4.3, 14.7, 24.2)	(15.7, 35.5, 48.3)	(42.6, 66.4, 76.9)	(79.3, 91.9, 95.5)	(98.3, 99.6, 99.9)
$\widehat{P}_{\mathrm{A}}$	(1.1, 5.0, 10.0)	(4.3, 14.8, 24.3)	(15.8, 35.7, 48.4)	(42.7, 66.5, 77.0)	(79.5, 92.0, 95.5)	(98.3, 99.6, 99.9)

Table S2. *Comparison of $p-$values, normal example, $n = 5, 10, 20$, nominal sizes $\alpha = (1, 5, 10)\%$, $\Delta = 0.05$, $g = 5$. Figures based on 50,000 replications, with 20,000 samples being drawn in calculation of each bootstrap $p-$value for $n = 5, 10$; 10,000 replications with 10,000 samples for each $p-$value for $n = 20$. Figures give percentages of p-values $< \alpha$. Testing against $H_a : \theta < \theta_0$.*

$\theta_0 =$	θ	$\theta + \Delta$	$\theta + 2\Delta$	$\theta + 3\Delta$	$\theta + 4\Delta$	$\theta + 5\Delta$
			$n = 5$			
$\widehat{P}_{\rm N}$	(7.7, 21.7, 33.1)	(10.6, 27.9, 40.5)	(14.1, 34.4, 48.0)	(18.1, 40.9, 55.5)	(22.7, 47.6, 62.5)	(27.6, 54.2, 68.8)
$\widehat{P}_{\rm B}$	(0.9, 4.8, 9.8)	(1.4, 6.9, 13.3)	(2.1, 9.5, 17.6)	(3.0, 12.5, 22.4)	(4.2, 16.1, 27.6)	(5.5, 19.9, 33.1)
$\widehat{P}_{\rm cB}$	(1.0, 5.2, 10.3)	(1.6, 7.5, 14.1)	(2.4, 10.2, 18.5)	(3.3, 13.4, 23.6)	(4.7, 17.3, 29.1)	(6.2, 21.4, 34.8)
$\widehat{P}_{\rm A}$	(1.1, 5.5, 11.0)	(1.7, 7.9, 14.9)	(2.6, 10.9, 19.5)	(3.6, 14.2, 24.7)	(5.0, 18.2, 30.2)	(6.7, 22.5, 35.9)
			$n = 10$			
$\widehat{P}_{\rm N}$	(4.1, 14.2, 23.7)	(7.3, 22.0, 34.2)	(11.9, 31.4, 45.4)	(17.9, 41.6, 56.7)	(25.2, 52.1, 66.9)	(33.3, 61.9, 75.7)
$\widehat{P}_{\rm B}$	(1.0, 4.8, 9.7)	(1.9, 8.5, 15.9)	(3.5, 13.7, 23.7)	(5.9, 20.2, 32.9)	(9.3, 28.1, 42.7)	(13.8, 36.8, 52.8)
$\widehat{P}_{\rm cB}$	(1.0, 5.0, 10.0)	(2.0, 8.8, 16.3)	(3.6, 14.1, 24.2)	(6.2, 20.8, 33.5))	(9.7, 28.8, 43.5)	(14.4, 37.7, 53.6)
$\widehat{P}_{\rm A}$	(1.0, 5.1, 10.2)	(2.1, 8.9, 16.6)	(3.8, 14.3, 24.50	(6.3, 21.1, 33.9)	(9.9, 29.2, 43.9)	(14.7, 38.1, 54.0)
			$n = 20$			
$\widehat{P}_{\rm N}$	(2.8, 10.7, 18.4)	(6.5, 20.8, 32.2)	(13.9, 34.4, 48.9)	(24.1, 50.1, 65.1)	(36.7, 65.4, 78.4)	(50.9, 78.1, 87.9)
$\widehat{P}_{\rm B}$	(1.0, 4.8, 10.1)	(3.0, 11.4, 19.9)	(6.4, 21.1, 33.1)	(13.2, 33.9, 48.9)	(22.4, 48.8, 64.5)	(33.7, 63.5, 77.3)
$\widehat{P}_{\rm cB}$	(1.0, 4.8, 10.1)	(3.0, 11.4, 19.9)	(6.4, 21.2, 33.2)	(13.3, 34.0, 49.0)	(22.4, 48.9, 64.6)	(33.8, 63.6, 77.4)
$\widehat{P}_{\rm A}$	(1.0, 4.9, 10.1)	(3.0, 11.5, 20.1)	(6.4, 21.3, 33.3)	(13.3, 34.2, 49.2)	(22.6, 49.1, 64.6)	(34.2, 63.8, 77.4)

Table S3. *Comparison of p−values, normal example,* $n = 5, 10, 15$, *nominal sizes* $\alpha = (1, 5, 10)\%$, $\Delta = 0.05$, $g = 10$. *Figures based on 10,000 replications, with 10,000 samples being drawn in calculation of each bootstrap p−value for* $n = 5, 10, 15$. *Figures give percentages of p-values* $< \alpha$. *Testing against* $H_a : \theta < \theta_0$.

$\theta_0 =$	θ	$\theta + \Delta$	$\theta + 2\Delta$	$\theta + 3\Delta$	$\theta + 4\Delta$	$\theta + 5\Delta$
			$n = 5$			
$\widehat{P}_{\mathrm{N}}$	(12.1, 29.9, 43.0)	(18.9, 41.0, 55.0)	(26.4, 52.2, 65.6)	(35.7, 62.2, 75.4)	(45.3, 71.5, 83.0)	(54.5, 79.4, 88.8)
$\widehat{P}_{\mathrm{B}}$	(1.0, 5.0, 9.6)	(1.9, 8.2, 15.6)	(3.5, 12.7, 22.4)	(5.5, 18.8, 30.6)	(8.2, 25.6, 39.9)	(12.0, 33.7, 49.3)
$\widehat{P}_{\mathrm{cB}}$	(1.1, 5.1, 10.0)	(2.1, 8.6, 16.2)	(3.7, 13.3, 23.1)	(5.9, 19.7, 31.6)	(8.8, 26.5, 41.0)	(12.7, 34.8, 50.6)
$\widehat{P}_{\mathrm{A}}$	(1.2, 5.7, 10.8)	(2.3, 9.2, 17.3)	(4.1, 14.4, 24.5)	(6.5, 21.0, 33.4)	(9.5, 28.1, 42.7)	(13.8, 36.6, 52.3)
			$n = 10$			
$\widehat{P}_{\mathrm{N}}$	(6.1, 19.2, 29.9)	(13.2, 33.0, 46.7)	(23.9, 49.0, 63.9)	(37.1, 65.2, 77.9)	(51.8, 78.3, 87.7)	(66.0, 87.4, 93.9)
$\widehat{P}_{\mathrm{B}}$	(0.9, 4.9, 9.8)	(2.6, 10.9, 19.7)	(6.2, 20.7, 32.8)	(12.2, 33.0, 47.8)	(21.4, 47.4, 63.0)	(32.4, 61.7, 76.0)
$\widehat{P}_{\mathrm{cB}}$	(1.0, 5.0, 9.9)	(2.7, 11.2, 20.0)	(6.3, 21.1, 33.1)	(12.6, 33.5, 48.3)	(21.9, 47.9, 63.5)	(33.1, 62.3, 76.4)
$\widehat{P}_{\mathrm{A}}$	(1.0, 5.2, 10.2)	(2.8, 11.4, 20.4)	(6.5, 21.5, 33.6)	(12.9, 34.0, 48.8)	(22.4, 48.4, 64.1)	(33.6, 62.8, 76.9)
			$n = 15$			
$\widehat{P}_{\mathrm{N}}$	(4.3, 15.1, 25.5)	(11.9, 32.4, 46.3)	(26.1, 52.8, 67.3)	(44.5, 71.7, 83.0)	(62.9, 85.6, 93.0)	(78.3, 93.8, 97.5)
$\widehat{P}_{\mathrm{B}}$	(1.1, 4.8, 9.4)	(3.4, 13.1, 23.4)	(9.1, 28.0, 42.3)	(20.9, 46.6, 61.8)	(36.3, 65.1, 78.2)	(53.9, 79.9, 89.8)
$\widehat{P}_{\mathrm{cB}}$	(1.1, 4.8, 9.5)	(3.5, 13.3, 23.6)	(9.3, 28.3, 42.5)	(21.0, 47.0, 62.1)	(36.6, 65.5, 78.3)	(54.4, 80.2, 89.9)
$\widehat{P}_{\mathrm{A}}$	(1.1, 4.9, 9.6)	(3.6, 13.5, 23.8)	(9.4, 28.5, 42.8)	(21.5, 47.2, 62.3)	(37.0, 65.9, 78.6)	(54.6, 80.4, 90.1)

Table S4. *Comparison of $p-$values, normal example, $n = 5, 10, 20$, nominal sizes $\alpha = (1, 5, 10)\%$, $\Delta = 0.05$, $g = 10$. All figures based on 10,000 replications, with 10,000 samples being drawn in calculation of each bootstrap $p-$value. Figures give percentages of p-values $< \alpha$. Testing against $H_a : \theta > \theta_0$.*

$\theta_0 =$	θ	$\theta - \Delta$	$\theta - 2\Delta$	$\theta - 3\Delta$	$\theta - 4\Delta$	$\theta - 5\Delta$
			$n = 5$			
$\widehat{P}_{\mathrm{N}}$	(0.1, 0.5, 1.1)	(0.2, 1.2, 2.8)	(0.7, 3.3, 6.4)	(2.1, 7.9, 13.7)	(6.1, 17.3, 26.4)	(15.4, 33.0, 44.0)
$\widehat{P}_{\mathrm{B}}$	(1.0, 4.8, 9.9)	(2.5, 9.9, 17.9)	(5.8, 18.5, 29.8)	(13.0, 32.0, 44.2)	(25.6, 47.9, 60.7)	(43.1, 65.9, 76.6)
$\widehat{P}_{\mathrm{cB}}$	(1.0, 4.9, 10.0)	(2.6, 10.1, 18.0)	(6.0, 18.8, 29.9)	(13.4, 32.1, 44.3)	(25.9, 48.0, 60.8)	(43.4, 66.0, 76.7)
$\widehat{P}_{\mathrm{A}}$	(1.0, 4.5, 9.3)	(2.4, 9.4, 17.1)	(5.7, 17.9, 28.6)	(12.8, 31.0, 43.0)	(24.8, 46.7, 59.4)	(42.2, 64.6, 75.6)
			$n = 10$			
$\widehat{P}_{\mathrm{N}}$	(0.1, 0.8, 2.3)	(0.6, 3.5, 7.3)	(3.0, 11.0, 18.5)	(10.8, 26.5, 38.5)	(28.1, 51.0, 63.5)	(55.5, 75.9, 84.2)
$\widehat{P}_{\mathrm{B}}$	(0.8, 5.0, 10.2)	(3.6, 13.3, 22.2)	(11.4, 28.9, 41.9)	(27.3, 51.5, 64.3)	(52.1, 74.1, 83.4)	(76.4, 90.8, 94.9)
$\widehat{P}_{\mathrm{cB}}$	(0.9, 5.0, 10.2)	(3.7, 13.4, 22.3)	(11.5, 29.1, 41.9)	(27.5, 51.6, 64.4)	(52.3, 74.2, 83.4)	(76.6, 90.9, 94.9)
$\widehat{P}_{\mathrm{A}}$	(0.9, 4.9, 10.0)	(3.6, 13.2, 22.0)	(11.3, 28.5, 41.4)	(27.0, 51.2, 64.0)	(51.8, 73.9, 83.1)	(76.5, 90.7, 94.7)
			$n = 20$			
$\widehat{P}_{\mathrm{N}}$	(0.2, 1.3, 2.9)	(1.3, 5.8, 10.9)	(6.4, 19.2, 29.6)	(22.4, 44.2, 57.3)	(51.6, 74.1, 83.5)	(81.7, 93.5, 96.5)
$\widehat{P}_{\mathrm{B}}$	(1.0, 4.8, 9.4)	(4.7, 15.5, 25.6)	(16.5, 37.4, 50.8)	(40.5, 65.2, 76.7)	(71.0, 88.4, 93.6)	(92.3, 97.7, 98.9)
$\widehat{P}_{\mathrm{cB}}$	(1.0, 4.9, 9.5)	(4.7, 15.6, 25.7)	(16.6, 37.5, 50.8)	(40.7, 65.3, 76.7)	(71.1, 88.5, 93.6)	(92.4, 97.7, 98.9)
$\widehat{P}_{\mathrm{A}}$	(1.0, 4.8, 9.4)	(4.6, 15.5, 25.5)	(16.5, 37.2, 50.6)	(40.5, 65.1, 76.6)	(70.8, 88.3, 93.6)	(92.4, 97.6, 98.9)

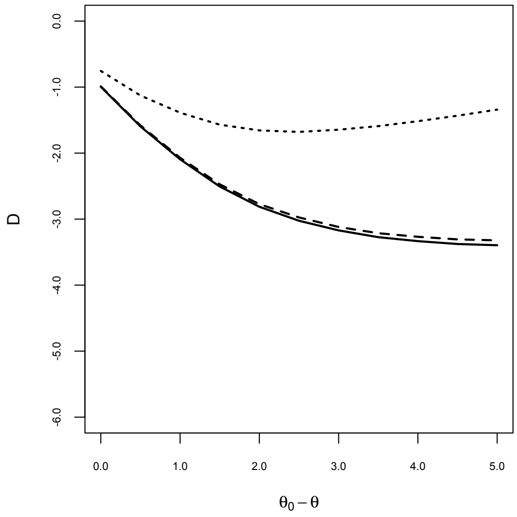

Fig. S1. Discrepancy D, inverse Gaussian example, $n = 5$, nominal size $\alpha = 5\%$, $\widehat{P}_{\mathrm{cB}}$ (solid), $\widehat{P}_{\mathrm{B}}$ (dots), $\widehat{P}_{\mathrm{A}}$ (dashes), testing $H_0 : \theta = \theta_0$ against $H_a : \theta < \theta_0$. Wald statistic.