Submitted to Statistical Science

A Diagnostic Function for the Accuracy of Bootstrap Confidence Intervals

Bradley Efron

Abstract. The standard intervals, a point estimate plus or minus some multiple of standard error, are mainstays of statistical practice. They are easy to calculate in most situations, while providing first-order asymptotic accuracy, but can be quite inaccurate in realistic applications. The bca system of bootstrap confidence intervals offers second-order accuracy, often a substantial improvement. These intervals are based on an implied transformation to a normal translation family. Is this justified? This paper implements a diagnostic function to answer the question. In any one application it produces a data-based estimate of accuracy for the bca intervals. The method applies to both parametric and nonparametric situations, and is particularly easy to use in the latter, requiring no special programming from the user.

Key words and phrases: equivalence, transformations to normality, correctness, nonparametric.

1. INTRODUCTION

Exact confidence intervals are an unattainable luxury in usual practice. Most often, statisticians use the *standard intervals* as an approximation,

$$\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}, \tag{1.1}$$

where $\hat{\theta}$ is an unbiased or nearly unbiased point estimate for a target parameter θ, $\hat{\sigma}$ is an estimate of its standard error, and $z^{(\alpha)}$ is the α percentile point of the standard normal distribution, cdf Φ,

$$z^{(\alpha)} = \Phi^{-1}(\alpha); \tag{1.2}$$

the familiar choice $\alpha = 0.975$, $z^{(\alpha)} = 1.96$, gives approximate 95% two-sided coverage, with 2.5% noncoverage on each side of (1.1).

The standard intervals enjoy two favorable properties that make them immensely popular:

1. They are asymptotically accurate, with actual coverage approaching the nominal level with errors of order $O(n^{-1/2})$ in sample size n (so-called "first order accuracy").
2. They are *automatic* in the sense of not requiring special theoretical calculations for each new application. This is especially true now that methods like the bootstrap painlessly provide estimates of the standard error $\hat{\sigma}$ in (1.1).

Department of Statistics, Sequoia Hall, Stanford University (e-mail: efron@stanford.edu).

The trouble with the standard intervals is that they can be quite inaccurate, as the examples that follow will show. The *bca* bootstrap confidence intervals, described in Section 2, are computer-intensive extensions of (1.1) that improve accuracy to second order, that is, to coverage errors of order $O(n^{-1})$ rather than $O(n^{-1/2})$, often a substantial practical improvement as well as a theoretical one.

The algorithm for the bca confidence limits is more complicated than (1.1), and involves Monte Carlo estimation of some adjustable constants. A reasonable concern is for the validity of the bca formula in one's own application. The *diagnostic function* of my title is an algorithm for checking the assumptions underlying the bca procedure.

Table 1 presents a small data set I will use for illustration: $n = 22$ students have each taken five tests with scores as shown (extracted from a larger set of 88 students in Mardia, Kent and Bibby, 1979). As a first example, suppose we are interested in the parameter

$$\theta = \text{correlation (mechanics score, vectors score)}. \tag{1.3}$$

Pearson's sample correlation coefficient $\hat{\theta}$ is the MLE of θ,

$$\hat{\theta} = 0.498, \tag{1.4}$$

assuming bivariate normality of the (mechanics, vectors) pairs, and independence across students. The time-tested standard error estimate for $\hat{\theta}$ is

$$\hat{\sigma} = \frac{1-\hat{\theta}^2}{(n-3)^{1/2}} = 0.173, \tag{1.5}$$

TABLE 1
Scores received by 22 students who have each taken tests in mechanics, vectors, algebra, analysis, and statistics. From Mardia, Kent and Bibby (1979).

	Mech	Vecs	Alg	Analy	Stat
1	7	51	43	17	22
2	44	69	53	53	53
3	49	41	61	49	64
4	59	70	68	62	56
5	34	42	50	47	29
6	46	40	47	29	17
7	0	40	21	9	14
8	32	45	49	57	64
9	49	57	47	39	26
10	52	64	60	63	54
11	44	61	52	62	46
12	36	59	51	45	51
13	42	60	54	49	33
14	5	30	44	36	18
15	22	58	53	56	41
16	18	51	40	56	30
17	41	63	49	46	34
18	48	38	41	44	33
19	31	42	48	54	68
20	42	69	61	55	45
21	46	49	53	59	37
22	63	63	65	70	63

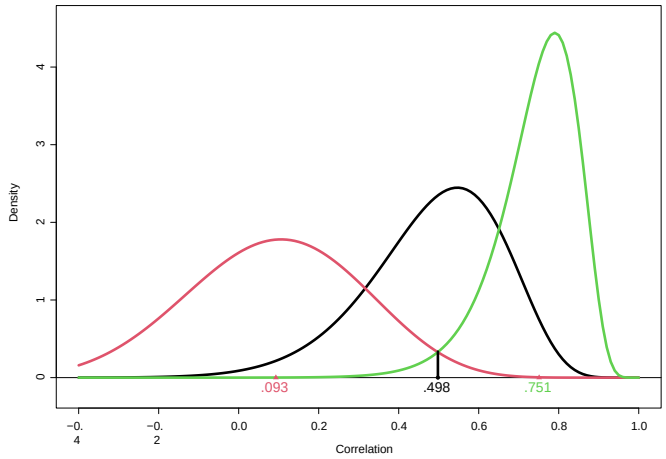

FIG 1. *Exact 95% interval for student score correlation is* $(0.093, 0.751)$*; densities* $f_\theta(\hat\theta)$ *for* $\theta = 0.093, 0.498, 0.751$.

yielding the 95% standard interval (1.1),

$$\theta \in [0.160, 0.836]. \tag{1.6}$$

This turns out to be quite wrong. The exact 95% interval is

$$\theta \in [0.093, 0.751]. \tag{1.7}$$

Figure 1 shows the "Neyman construction" of the exact interval: the choice $\theta = 0.093$ gives $\hat\theta$ probability 0.025 of exceeding the observed value 0.498, while $\theta = 0.751$ gives probability 0.025 of $\hat\theta$ being less than 0.498. The three curves show the density function $f_\theta(\hat\theta)$ for $\theta = 0.093, 0.498$, and 0.751.

The standard interval (1.1) takes literally the asymptotic normal model

$$\hat\theta \sim \mathcal{N}(\theta, \sigma^2), \tag{1.8}$$

a *normal translation family* (ntf). The actual family of densities $f_\theta(\hat\theta)$ for the normal theory estimates $\hat\theta$, given true correlation θ, is emphatically not ntf, as Figure 1 shows, foretelling the poor performance of (1.1).

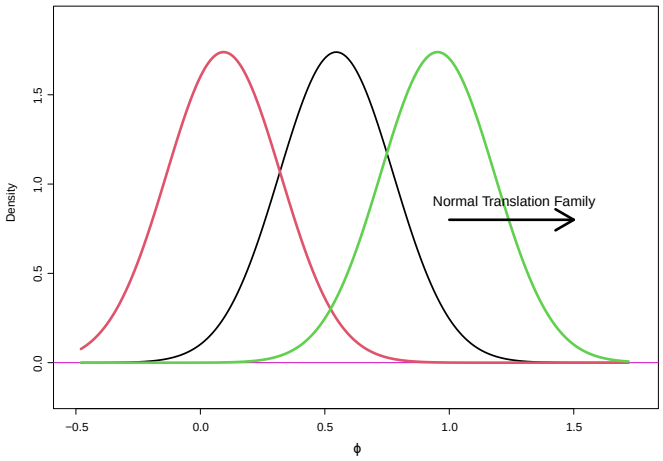

FIG 2. *The three density curves from Figure 1 now transformed to Fisher's* ϕ *scale.*

Fisher suggested a transformation of variables to put the correlation coefficient family $f_\theta(\hat\theta)$ into ntf form:

$$\phi = \frac{1}{2}\log\left(\frac{1+\theta}{1-\theta}\right) \quad \text{and} \quad \hat\phi = \frac{1}{2}\log\left(\frac{1+\hat\theta}{1-\hat\theta}\right). \tag{1.9}$$

Figure 2 shows the three densities from Figure 1, now for the transformed family $f_{\hat\phi}(\hat\phi)$. To the eye at least, the ntf ideal $\hat\phi \sim \mathcal{N}(\phi, \sigma^2)$ seems to have been achieved (with $\sigma = (n-3)^{-1/2} = 0.229$). The standard interval on the ϕ scale, now fully justified,[1] is

$$\hat\phi \pm 1.96\sigma. \tag{1.10}$$

The inverse mapping of ϕ to θ applied to (1.10) gives a quite accurate 95% interval for θ, as shown in Table 2.

Fisher's method is brilliantly ingenious, and suited to the computational limitations of 1915, but it has two points against it:

- *It requires Fisher.* Given a new situation, the statistician needs to figure out what mapping, if any, would play the role of (1.9) in transforming $f_\theta(\hat\theta)$ to a normal translation family–which is to say that the transformation method is not automatic.

[1] In fact, $f_{\hat\phi}(\hat\phi)$ is not *exactly* $\mathcal{N}(\phi, \sigma^2)$, as we'll see.

TABLE 2
Approximate and exact 95% confidence limits for student score correlation coefficient (1.3)*; bca is bootstrap interval described in Section 2, which in this case can be calculated directly, without simulation.*

	.025	.975
Standard	.160	.836
Fisher	.097	.760
Exact	**.093**	**.751**
bca	.083	.754

TABLE 3
Standard and bca approximate 95% confidence limits for student score maximum eigenvalue (1.12)*.*

	.025	.975
Standard	288	1078
bca	424	1426

- *It can't deal with nuisance parameters.* There are none in the bivariate normal correlation coefficient problem, but that is unusual even in small-sample situations.

The standard method (1.1) overcomes both of these limitations, at the expense of possibly serious inaccuracies. The bottom line in Table 2, "bca", refers to the bootstrap confidence interval algorithm described in Section 2. It also has an automatic character, with a single algorithm applying to a wide range of situations, including those with nuisance parameters (as well as to nonparametric settings; see Section 5), and with higher accuracy than the standard method, already evident in Table 2.

The example in Table 3 better illustrates the practical difficulties of setting confidence intervals. We assume, as in Mardia, Kent and Bibby (1979), that the rows of the student score data matrix in Table 1 are independent draws from a five-dimensional normal distribution,

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}_5(\gamma, \Sigma), \quad i = 1, \ldots, 22, \tag{1.11}$$

(γ a 5-vector and Σ a 5×5 matrix, both unknown) and that the parameter of interest θ is the maximum eigenvalue "maxeig" of Σ,

$$\theta = \text{maxeig}(\Sigma). \tag{1.12}$$

The maximum likelihood estimate $\hat{\theta}$ is

$$\hat{\theta} = \text{maxeig}\left(\widehat{\Sigma}\right) = 683, \tag{1.13}$$

where $\widehat{\Sigma}$ is the MLE of Σ ($21/22$ times the usual unbiased estimate). We want confidence limits for θ.

Table 3 shows the lower and upper limits for the standard and bca approximate 95% confidence intervals for θ. The bca interval is much shorter to the left of $\hat{\theta} = 683$, and much longer to the right. Can we believe it? In this case there is no gold standard exact interval for comparison. This is where the diagnostic function comes in. I will show (in Section 5) that in this situation the assumptions underlying the bca interval are well-supported by the data.

The paper proceeds as follows:

- Section 2 reviews the bca system of bootstrap confidence intervals and the distributional assumptions ("NSTF") that support it.
- A diagnostic function is introduced in Section 3 that is linear in NSTF families but nonlinear in a wider class ("GSTF") of distributional families.
- A GSTF version of the bca intervals ("gbca") is developed in Section 4, leading to *equivalence tables* that show the accuracy of the bca intervals if gbca is actually the appropriate choice.
- Section 5 extends the previous results to multiparameter families, including nuisance parameters, and develops a computational procedure for carrying out the diagnostic/equivalence calculations.
- A collection of remarks and details is presented in Section 6, which ends with a brief discussion.

2. THE BCA SYSTEM OF BOOTSTRAP CONFIDENCE INTERVALS

The standard intervals (1.1) have as their *target class* the normal translation families (ntf), $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$. If the distribution of $\hat{\theta}$ given θ is an ntf then $\hat{\theta} \pm z^{(\alpha)}\sigma$ is a perfectly accurate confidence interval. One way to say what goes wrong in our previous example is that ntf is a small and specialized target class, and one that's easy to miss.

The bca system of confidence intervals aims at a bigger target class: normal scaled transformation families, or NSTF (Efron, 1982). A family of distributions of $\hat{\theta}$ given θ is NSTF if there exists a monotone transformation $m(\cdot)$ such that

$$\phi = m(\theta) \text{ and } \hat{\phi} = m\big(\hat{\theta}\big) \tag{2.1}$$

satisfy

$$\hat{\phi} = \phi + (1 + a_0\phi)(Z - z_0), \tag{2.2}$$

with $Z \sim \mathcal{N}(0, 1)$. Here z_0 is the "bias corrector" and a_0 the "acceleration", as discussed next. Fisher's intervals for the correlation coefficient assumed that $\hat{\theta}$ given θ was an NSTF, with transformation $m(\cdot)$ as in (1.9), and z_0 and a_0 both zero. Section 3 will show that to be nearly, but not exactly, correct.

The NSTF class expands on the normal translation model $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$ in three ways:

1. By allowing for a monotone transformation $m(\cdot)$ to a normal scale (2.2).
2. By allowing for a bias correction z_0 on the normalized scale.

3. By allowing for non-constant standard deviation on the normalized scale, as measured by the acceleration a_0.

All three corrections to the ntf model are necessary for second order accuracy; see Hall (1988) and Section 8 of DiCiccio and Efron (1996).

Accuracy — the agreement of nominal and actual coverage rates — is by itself not a sufficient performance criterion. Fisher argued that his transformation argument produced *correct* as well as accurate intervals for the correlation problem: given $\hat{\phi} \sim \mathcal{N}(\phi, \sigma^2)$, the interval $\hat{\phi} \pm 1.96\sigma$ uses all the available information, in a frequentistically obvious way. If I randomly chose 11 of the 22 students in Table 1 and used their data to form an exact interval for θ, it would be perfectly accurate but incorrect. Correctness is one of those Fisherian ideas that's hard to pin down but important to consider.

Let $F_\theta(\hat{\theta})$ be the cdf of observation $\hat{\theta}$ given parameter value θ. The bca confidence limits depend on the following theorem from Efron (1987):

THEOREM 1. *If the family of cdfs $F_\theta(\hat{\theta})$ is an NSTF then the exact and correct one-sided level α upper confidence limit $\theta_{\rm bca}[\alpha]$ for θ is*

$$\hat{\theta}_{\rm bca}[\alpha] = F_{\hat{\theta}}^{-1}\left\{\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a_0(z_0 + z^{(\alpha)})}\right)\right\}, \tag{2.3}$$

with Φ the standard normal cdf and $z^{(\alpha)} = \Phi^{-1}(\alpha)$ as before. The two-sided level $2\alpha - 1$ bca confidence interval is

$$\theta \in \big(\hat{\theta}_{\rm bca}[1-\alpha], \hat{\theta}_{\rm bca}[\alpha]\big), \tag{2.4}$$

and this will give exactly the claimed coverage probabilities in an NSTF.

Notice that the monotone function $m(\theta)$ in (2.1)–(2.2) is *not* required to calculate (2.3); only its existence is needed.

Theorem 1 can be re-expressed in bootstrap terms. Let

$$\beta(\alpha) = \Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a_0(z_0 + z^{(\alpha)})}\right), \tag{2.5}$$

so $\hat{\theta}_{\rm bca}[\alpha] = F_{\hat{\theta}}^{-1}(\beta(\alpha))$. But $F_{\hat{\theta}}$ is the cdf of the *bootstrap distribution* of observations $\hat{\theta}^*$, given parameter value $\theta = \hat{\theta}$. (The star notation is needed to differentiate $\hat{\theta}^*$ from the original observations $\hat{\theta}$.) Therefore $F_\theta^{-1}(\beta)$ is the β level bootstrap percentile of $\hat{\theta}^*$, say $\hat{\theta}^{*(\beta)}$, giving bca confidence limit

$$\hat{\theta}_{\rm bca}[\alpha] = \hat{\theta}^{*(\beta(\alpha))}, \tag{2.6}$$

the $\beta(\alpha)$th bootstrap percentile.

A first guess at forming bootstrap confidence intervals might take $\beta(\alpha)$ equal α, which amounts to using the "percentile method"; that is, to setting $\hat{\theta}[\alpha]$ equal the αth bootstrap percentile. This is what we get from (2.3) if z_0 and a_0 both equal zero. They both do *go* to zero at stochastic rate $O_p(n^{-1/2})$ in sample size n. In practice, however, the corrections $\beta(\alpha)$ in (2.3) are necessary for the accuracy of bca limits.

The bootstrap cdf $\widehat{F} = F_{\hat{\theta}}$ usually needs to be approximated by bootstrap resampling, either parametric or nonparametric, as illustrated by the examples in Section 4 and Section 5. $\widehat{F}$ also provides an estimate of the bias corrector z_0,

$$z_0 = \Phi^{-1}(p_0) \text{ where } p_0 = \widehat{F}\big(\hat{\theta}\big). \tag{2.7}$$

If $p_0 < 0.50$, that is, if less than 50% of the bootstrap replications $\hat{\theta}^*$ are less than $\hat{\theta}$, then $\hat{\theta}^*$ is upwardly biased and the bias correction z_0 is in the negative direction; and conversely for $p_0 > 0.50$.

For the student score correlation coefficient with $\hat{\theta} = 0.498$,

$$p_0 = \Pr_{\hat{\theta}}\big\{\hat{\theta}^* \le \hat{\theta}\big\} = 0.478 \tag{2.8}$$

and $z_0 = -0.056$, being a small negative bias correction. Fisher's method assumes $z_0 = 0$, partly accounting for its discrepancies from the exact limits in Table 2. Section 3 reveals another cause.

By itself, the bootstrap distribution $\widehat{F} = F_{\hat{\theta}}$ cannot specify a confidence interval for θ. We need to know something about distributions F_θ for θ away from $\hat{\theta}$. In Figure 1 the central curve represents the parametric bootstrap distributions, but it's the two outlying curves that determine the exact confidence limits.

The acceleration parameter a_0 says just enough about deviations of F_θ from $F_{\hat{\theta}}$ to allow construction of the bca limits. Equation (4.4) of Efron (1987) gives the formula for a_0 in terms of the score function of the family F_θ,

$$a_0 = \frac{1}{6}\text{skewness (score function)}. \tag{2.9}$$

For the normal correlation coefficient, a_0 was very close to zero.

The bca confidence limits have several attractive features:

- They are *transformation respecting*: in notation (2.1),
  $$\hat{\phi}_{\rm bca}[\alpha] = m\big(\hat{\theta}_{\rm bca}[\alpha]\big), \tag{2.10}$$
  so there is never a question of working on the wrong scale.
- They are *second order accurate*, with coverage errors approaching nominal values at rate $O(n^{-1})$ in sample size n, which can make a major difference in examples like that of Table 3.
- The are *correct* as well as accurate, in the same sense as Fisher's method for the normal correlation coefficient; see Section 3.

- They are *automatable*: a single program is available to handle a wide variety of situations without requiring special work on the user's part. See Remark 13; this is especially convenient in nonparametric applications.

In any one situation, the actual performance of the bca intervals depends on how well the NSTF assumptions (2.1)–(2.2) apply to the problem at hand. The diagnostic function discussed next offers a data-based check on the assumptions.

3. THE DIAGNOSTIC FUNCTION

The bca confidence interval endpoints (2.3) depend for their validity on the NSTF assumptions (2.1)–(2.2): that there exists a monotone transformation $\phi = m(\theta)$ and $\hat{\phi} = m(\hat{\theta})$ that results in a normal translation family (augmented to allow for bias and changing standard error). The diagnostic function of my title, discussed next, permits us to test for NSTF structure against a wider class of possibilities.

We have a one-parameter family of distributions denoted in cdf form as

$$\left\{F_\theta\big(\hat{\theta}\big),\ \theta \in \Theta\right\}, \tag{3.1}$$

for Θ an interval of the real line; (3.1) is a *generalized scaled transformation family* (GSTF) if there exists a monotone transformation $\phi = m(\theta)$ and $\hat{\phi} = m(\hat{\theta})$ such that

$$\begin{aligned} &\hat{\phi} = \phi + (1 + a_0\phi)\big(\widetilde{Z} - \tilde{z}_0\big), \\ &\text{where} \quad \widetilde{Z} = w(Z) \quad [Z \sim \mathcal{N}(0,1)], \end{aligned} \tag{3.2}$$

with $w(\cdot)$ a differentiable monotone increasing function satisfying

$$w(0) = 0 \ \text{ and } \ w'(0) = 1; \tag{3.3}$$

also, from (2.7),

$$\tilde{z}_0 = w(z_0). \tag{3.4}$$

A GSTF is NSTF if $w(z) = z$. See Remark 1 in Section 6.

GSTF generalizes NSTF by allowing non-normality on the ϕ scale. It will turn out, for example, that the normal correlation family is better described with $\widetilde{Z}$ a student-t distribution rather than $\mathcal{N}(0,1)$. For $n = 22$ and $\theta = 0.498$, the degrees of freedom ν for the student-t distribution is about $\nu = 63$, so

$$w(z) = T_{63}^{-1}(\Phi(z)), \tag{3.5}$$

where T_{63} is the student-t cdf; to the eye, $w(z)$ is almost indistinguishable from z.

The diagnostic function $D_\theta(z)$ was introduced in Efron (1982) as an answer to the question of whether a one-parameter family of distributions (3.1) could be transformed into a normal translation family. For

$$\alpha = \Phi^{-1}(z), \tag{3.6}$$

we define

$$C_\theta(z) = \frac{\dot{F}_\theta\big(\hat{\theta}_\theta^{(\alpha)}\big)}{\varphi(z)}, \tag{3.7}$$

where $\dot{F}_\theta(\hat{\theta}) = (\partial/\partial\theta)F_\theta(\hat{\theta})$, $\hat{\theta}_\theta^{(\alpha)}$ is the αth percentile of $\hat{\theta}$ given θ,

$$F_\theta\big(\hat{\theta}_\theta^{(\alpha)}\big) = \alpha, \tag{3.8}$$

finally giving the diagnostic function $D_\theta(z)$,

$$D_\theta(z) = C_\theta(z)\big/C_\theta(0). \tag{3.9}$$

$D_\theta(z)$ is *transformation invariant*, remaining unchanged under transformations such as (2.1).

Section 2 of Efron (1982) proves the following theorem:

THEOREM 2. *The diagnostic function for a GSTF, that is a family satisfying* (3.2)–(3.4)*, is*

$$D_\theta(z) = \frac{1 + w(z)\epsilon_0}{w'(z)}, \tag{3.10}$$

where

$$\epsilon_0 = \frac{a_0}{1 - a_0\tilde{z}_0}. \tag{3.11}$$

Both a_0 and $\tilde{z}_0$ are usually small, making $\epsilon_0 \doteq a_0$, the rate of change of standard deviation on the ϕ scale (Section 2).

In an NSTF (2.1)–(2.2), we have $w(z) = z$ and $w'(z) = 1$, giving this informative result:

COROLLARY (to Theorem 2). *The diagnostic function for an NSTF is*

$$D_\theta(z) = 1 + z\epsilon_0, \tag{3.12}$$

with $\epsilon_0 = a_0/(1 - a_0 z_0)$.

In other words, if $\{F_\theta(z), \theta \in \Theta\}$ is an NSTF (2.1)–(2.2), then $D_\theta(z)$ will be *linear* with slope ϵ_0. There is also a converse: if a_0 and z_0 are known (as they will be from the bootstrap algorithm of Section 5) then $\epsilon_0 = a_0/(1 - a_0 z_0)$, and $D_\theta(z) = 1 + z\epsilon_0$ implies that $\{F_\theta(\hat{\theta})\}$ is an NSTF. In this case Theorem 1 says that *if $D_\theta(z)$ is linear, then the bca confidence limits* (2.3) *are exact.* Remark 2 in Section 6 verifies the converse to the corollary.

As a first example, suppose $\{F_\theta(\hat{\theta})\}$ represents a *gamma scale family*,

$$\hat{\theta} = \theta G_\nu/\nu, \tag{3.13}$$

where G_ν is a standard gamma variate with ν degrees of freedom,

$$f_\theta\big(\hat{\theta}\big) = \frac{\hat{\theta}^{\nu-1} e^{-\nu\hat{\theta}/\theta}\left(\frac{\nu}{\theta}\right)^\nu}{\Gamma(\nu)}. \tag{3.14}$$

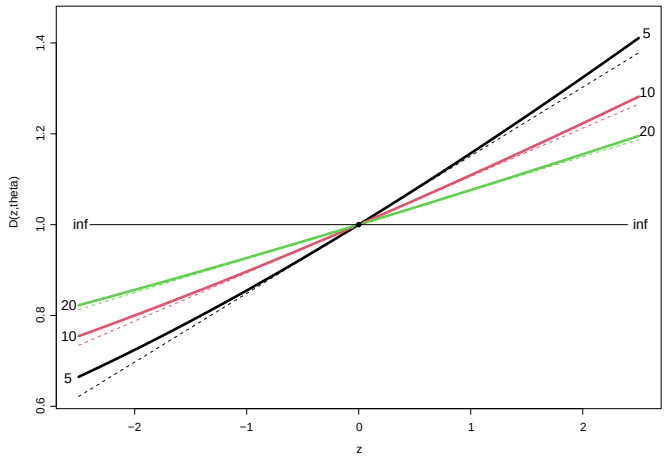

FIG 3. *Diagnostic function for Gamma scale family, df*$= 5, 10, 20$ *and infinity; dashed lines are linear fits.*

TABLE 4
Exact and bca limits for Gamma scale family, $\nu = 5$.

α	.025	.05	.1	.9	.95	.975
Exact	.4882	.5462	.6255	2.055	2.538	3.080
bca	.4884	.5464	.6256	2.056	2.541	3.089

Figure 3 shows $D_\theta(z)$ for four cases: $\nu = 5, 10, 20$ and the limiting case $\nu \to \infty$.[2] ($D_\theta(z)$ is the same for all θ in family (3.13).) The plots look nearly linear, with a slight upward bend, most evident for $\nu = 5$. As $\nu \to \infty$, the diagnostic function approaches

$$D_\theta(z) = 1 \quad \text{for all } z, \tag{3.15}$$

the diagnostic function for a normal translation family (1.8).

Is "nearly linear" good enough to guarantee the accuracy of the bca confidence intervals? The examples in what follows show generally good bca performance even under much more substantial deviations of $D_\theta(z)$ from linearity, but the method can be pushed too far; see Remark 5 in Section 6. For the gamma scale family (3.13) we can compare the bca limits $\hat\theta_{\text{bca}}[\alpha]$ with the exact limits

$$\hat\theta_{\text{exact}}[\alpha] = \nu\hat\theta \big/ G_\nu^{(1-\alpha)}, \tag{3.16}$$

with $G_\nu^{(1-\alpha)}$ the $(1-\alpha)$ percentile of G_ν. Table 4 makes the comparison for $\nu = 5$, $\hat\theta = 1$, where we see almost perfect agreement.

The slopes of the linear fits for $\nu = 5, 10, 20$ in Figure 3 are

$$\hat\epsilon_0 = (0.150, 0.106, 0.0745) \tag{3.17}$$

compared with the theoretical values (3.11)

$$\epsilon_0 = (0.152, 0.107, 0.0750) \tag{3.18}$$

obtained from (2.7) and (2.9).

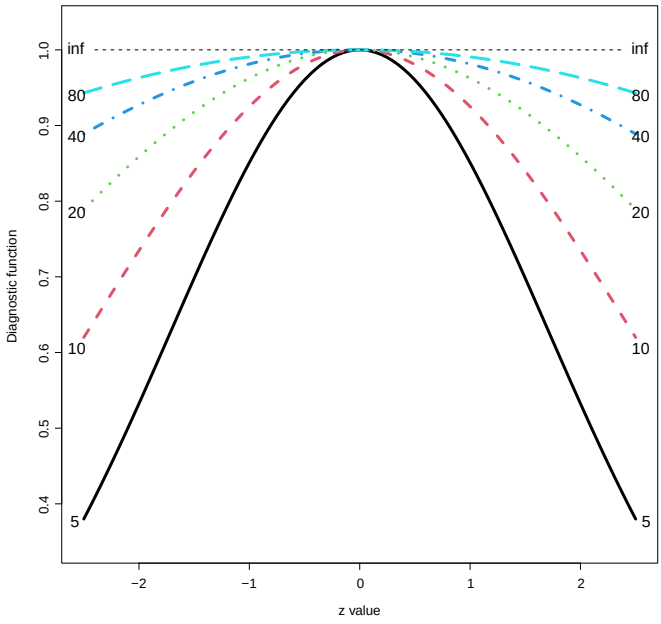

FIG 4. *Diagnostic function for student-t translation model; degrees of freedom 5, 10, 20, 40, 80 and infinity.*

A less reassuring example is pictured in Figure 4. Here

$$F_\theta\big(\hat\theta\big) = \Pr\big\{\widetilde{Z}_\nu \le \hat\theta - \theta\big\}, \tag{3.19}$$

where $\widetilde{Z}_\nu$ has a student-t distribution with ν degrees of freedom. That is,

$$\hat\theta = \theta + \widetilde{Z}_\nu, \tag{3.20}$$

a student-t translation model; this is a GSTF (3.2) with

$$\phi = \theta, \quad z_0 = a_0 = 0, \quad \text{and} \quad \widetilde{Z} = \widetilde{Z}_\nu. \tag{3.21}$$

Figure 4 shows diagnostic function $D_\theta(z)$ (the same for all θ) for degrees of freedom $\nu = 5, 10, 20, 40, 80$, and $\nu \to \infty$, the last being the ntf diagnostic $D_\theta(z) = 1$. $D_\theta(z)$ is definitely *nonlinear*, spectacularly so for $\nu = 5$, but this doesn't necessarily mean that the bca intervals are wrong. In fact, Remark 6 of Section 6 shows that in situation (3.21) formulat (2.3) is exactly correct. (That's not true if z_0 or a_0 is nonzero in (3.21).)

The heavy dashed curve in Figure 5 is the diagnostic function relating to the correlation example in Section 1. That is, $F_{\hat\theta}(\hat\theta^*)$ is the cdf for the Pearson correlation coefficient $\hat\theta^*$ from a sample of $n = 22$ bivariate normal pairs having true correlation $\hat\theta = 0.498$. (Here the diagnostic calculation (3.4)–(3.9) can be done theoretically, the bootstrap notation being necessary only to avoid using variable $\hat\theta$ for both true and observed correlations.)

[2] Remark 3 in Section 6 describes the calculation of $D_\theta(z)$.

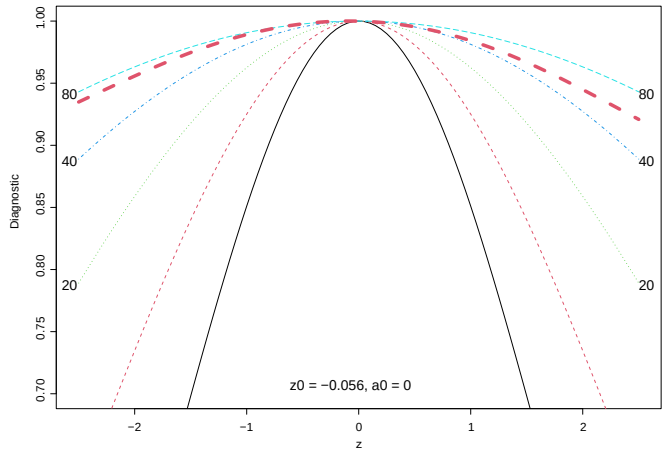

FIG 5. *Diagnostic function for normal correlation Coef,* $n = 22$ *(heavy dashed curve), compared to student-t translates from Figure 4.*

$D_\theta(z)$ is seen to be slightly concave, like the student-t translation diagnostic in Figure 4. The amount of curvature matches a student-t curve with 63 degrees of freedom, though with a slight downward tilt left to right. Fisher was nearly right, but not exactly so, in assuming transformation (1.9) produced a normal translation family.

4. GENERALIZED BOOTSTRAP CONFIDENCE INTERVALS

We have a one-parameter family of cdfs $\{F_\theta(\hat\theta), \theta \in \Theta\}$ and wish to set confidence limits for θ having observed $\hat\theta$. If the family satisfies the NSTF conditions (2.1)–(2.2), Theorem 1 says that the bca limits (2.3) are exactly accurate and correct. There is a generalized version of Theorem 1 that applies to GSTF families (3.2), discussed next.

Formula (2.3) says that the level α bca confidence limit $\hat\theta_{\rm bca}[\alpha]$ is the $\beta[\alpha]$ percentile of the bootstrap distribution $\widehat F = F_{\hat\theta}$, where $\beta[\alpha]$ is given by (2.5). We can write $\beta[\alpha]$ in terms of the standard normal cdf $\Phi(z)$:

$$\beta[\alpha] = \Phi\left\{\Phi^{-1}(p_0) + \frac{\Phi^{-1}(p_0)+\Phi^{-1}(\alpha)}{1-a_0(\Phi^{-1}(p_0)+\Phi^{-1}(\alpha))}\right\}, \tag{4.1}$$

$p_0 = \Phi(z_0) = \widehat F(\hat\theta)$ (2.7). The GSTF conditions (3.2)–(3.4) replace $Z \sim \mathcal{N}(0,1)$ with

$$\widetilde Z = w(Z). \tag{4.2}$$

Define $\widetilde\Phi$ to be the cdf of $\widetilde Z$,

$$\widetilde\Phi(z) = \Pr\{\widetilde Z \le z\} = \Phi(w^{-1}(z)), \tag{4.3}$$

and

$$\tilde\beta[\alpha] = \widetilde\Phi\left\{\widetilde\Phi^{-1}(p_0) + \frac{\widetilde\Phi^{-1}(p_0)-\widetilde\Phi^{-1}(1-\alpha)}{1-a_0(\widetilde\Phi^{-1}(p_0)-\widetilde\Phi^{-1}(1-\alpha))}\right\}. \tag{4.4}$$

THEOREM 3. *If* $\{F_\theta(\hat\theta), \theta \in \Theta\}$ *is a GSTF* (3.2)*, the exact and correct level* α *upper confidence limit* $\hat\theta_{\rm gbca}[\alpha]$ *for* θ *is*

$$\hat\theta_{\rm gbca}[\alpha] = \widehat F^{-1}\big(\tilde\beta[\alpha]\big). \tag{4.5}$$

The proof of Theorem 2 in Efron (1987) doesn't rely on any special properties of the normal cdf $\Phi(z)$ except for its symmetry, $z^{(\alpha)} = -z^{1-\alpha)}$ (expression (4.4) equals (4.1) if $\widetilde\Phi = \Phi$). The key step is to rewrite (3.2) as

$$(1+a_0\hat\phi) = (1+a_0\phi)\big(1+a_0(\widetilde Z - \tilde z_0)\big). \tag{4.6}$$

Taking logarithms gives

$$\begin{aligned} \hat\zeta &= \zeta + Q, \quad \text{where} \\ \hat\zeta &= \log(1+a_0\hat\phi), \\ \zeta &= \log(1+a_0\phi), \quad \text{and} \\ Q &= \log\big(1+a_0(\widetilde Z - \tilde z_0)\big). \end{aligned} \tag{4.7}$$

The original parameterization $\{F_\theta(\hat\theta)\}$ has now been transformed into a translation family $\hat\zeta = \zeta + Q$ according to the monotone mapping

$$q(\cdot) = \log(1+a_0 m(\cdot)), \tag{4.8}$$

$\phi = m(\theta)$ as before. The exact level α confidence limit in translation family (4.7) is

$$\hat\zeta[\alpha] = \hat\zeta - Q^{(1-\alpha)} \tag{4.9}$$

since

$$\Pr\{\zeta \le \hat\zeta[\alpha]\} = \Pr\{\hat\zeta \ge \hat\zeta^{(1-\alpha)}\} = \alpha. \tag{4.10}$$

By monotonicity, the inverse mapping $\hat\theta[\alpha] = q^{-1}(\hat\zeta[\alpha])$ is exactly accurate. This gives Theorem 3; see Remark 7 for the proof.

The confidence limit (4.9) is *correct* as well as accurate, in the Fisherian sense discussed in Section 1: model $\hat\zeta = \zeta + Q$ adds random variate Q to ζ, so (4.9) amounts to stochastic subtraction.

The bca confidence limit $\hat\theta_{\rm bca}[\alpha]$ depends on the NSTF model (2.1)–(2.2), while $\hat\theta_{\rm gbca}[\alpha]$ assumes the broader GSTF model (3.2). How well justified are the NSTF assumptions? Since $\hat\theta_{\rm bca}[\alpha] = \widehat F^{-1}(\beta[\alpha])$ while $\hat\theta_{\rm gbca}[\alpha] = \widehat F^{-1}(\tilde\beta[\alpha])$ (4.5), the question amounts to asking how closely does $\beta[\alpha]$ (4.1) match $\tilde\beta[\alpha]$ (4.4)? An answer is developed for one-parameter models in what follows and then extended to multiparameter models, with nuisance parameters, in Section 5.

Having observed $\hat\theta$ from model $F_\theta(\hat\theta)$ we obtain the bootstrap cdf $\widehat F = F_{\hat\theta}$, $p_0 = F_\theta(\hat\theta)$, a_0 (2.9) and the diagnostic function $D_{\hat\theta}(z)$, i.e., $D_\theta(z)$ (3.6)–(3.9) evaluated at $\theta = \hat\theta$ (called $D_0(z)$ below). $D_\theta(z)$ doesn't depend on θ in GSTF families; see Remark 8. None of these steps relies on the choice of NSTF versus GSTF; they also give $z_0 = \Phi^{-1}(p_0)$ and $\epsilon_0 = a_0/(1 - a_0 z_0)$ (3.11).

Given ϵ_0, the diagnostic function (3.10)

$$D_0(z) = \frac{1 + w(z)\epsilon_0}{w'(z)} \tag{4.11}$$

determines the function $w(z)$. Let

$$I(z) = \int_0^z \frac{1}{D_0(y)}\, dy. \tag{4.12}$$

THEOREM 4. *If $\epsilon_0 = 0$,*

$$w(z) = I(z). \tag{4.13}$$

Otherwise

$$w(z) = \frac{1}{\epsilon_0}(e^{\epsilon_0 I(z)} - 1). \tag{4.14}$$

PROOF. If $\epsilon_0 = 0$ then (4.13) follows from $w'(z) = D_0(z)^{-1} = I'(z)$, and $w(0) = I(0) = 0$ (3.3). If $\epsilon_0 \neq 0$, differentiating (4.14) gives

$$\begin{aligned} w'(z) &= I'(z)e^{\epsilon_0 I(z)} \\ &= I'(z)(1 + w(z)\epsilon_0) \\ &= \frac{1}{D(z)}(1 + w(z)\epsilon_0), \end{aligned} \tag{4.15}$$

which is (4.11). □

The function $w(z)$ determines the cdf $\widetilde\Phi(z)$ (4.3),

$$\widetilde\Phi(z) = \Pr\{w(Z) \le z\} = \Phi(w^{-1}(z)), \tag{4.16}$$

and also

$$\tilde z_0 = \widetilde\Phi^{-1}(p_0) = \widetilde Z^{(p_0)} = w(Z^{(p_0)}) = w(z_0). \tag{4.17}$$

(Because $w(0) = 0$ and $w'(0) = 1$, $\tilde z_0$ is usually close to z_0.)

We now have all the ingredients for $\tilde\beta[\alpha]$ (4.4). In the following examples an *equivalent level* $\tilde\alpha$ is calculated; that is, the value $\tilde\alpha$ that makes

$$\tilde\beta(\tilde\alpha) = \beta(\alpha). \tag{4.18}$$

If $\tilde\alpha$ is close to α then $\hat\theta_{\rm bca}[\alpha]$ is close to $\hat\theta_{\rm gbca}[\alpha]$, supporting the robustness of the bca method to deviations from the NSTF assumptions.

Table 5 shows the equivalent level $\tilde\alpha$ for six choices of α, for the student score normal correlation coefficient, Table 2. In this case the agreement is nearly perfect, though there is a hint of overcoverage at $\alpha = 0.025$ and undercoverage at $\alpha = 0.975$. (We also saw that in Table 2, where we had the advantage of knowing exact limits.) The examples of this paper show generally good accuracy for the bca intervals, but see Remark 9 for a counterexample.

TABLE 5
Equivalent GSTF level $\tilde\alpha$ for nominal NSTF level α (4.18)*; student score correlation coefficient (Table 2).*

α	.025	.05	.1	.9	.95	.975
$\tilde\alpha$	.0247	.0495	.0995	.8996	.9496	.9747

The corollary in Section 3 says that diagnostic function $D_\theta(z)$ is linear if and only if $\{F_\theta(\hat\theta)\}$ is NSTF. Theorem 5 describes the nonlinearity of $D_\theta(z)$ in GSTF models. We start with the GSTF model (3.2),

$$\hat\phi = \phi + \sigma_\phi(\widetilde Z - \tilde z_0) \quad (\sigma_\phi = 1 + a_0\phi), \tag{4.19}$$

where $\widetilde Z = w(Z)$ has cdf $\widetilde\Phi(\cdot)$, and $\tilde z_0 = \widetilde\Phi^{-1}(p_0)$ as in (4.2)–(4.4). For convenience we assume that $\widetilde Z$ has median zero

$$\tilde z^{(0.5)} = \widetilde\Phi^{-1}(0.5) = 0 \tag{4.20}$$

(which can always be achieved by subtracting the original median from $\widetilde Z$ and $\tilde z_0$).

THEOREM 5. *A GSTF* (4.19)–(4.20) *has diagnostic function*

$$D_0(z) = (1 + \hat z^{(\alpha)}\epsilon_0)\cdot r(z), \tag{4.21}$$

with $\epsilon_0 = a_0(1 - a_0\tilde z_0)^{-1}$, $\widetilde Z^{(\alpha)} = \widetilde\Phi^{-1}(\alpha)$, the α percentile of $\widetilde Z$, and

$$r(z) = \frac{\tilde\varphi(\hat z^{(\alpha)})/\tilde\varphi(0)}{\varphi(z^{(\alpha)})/\varphi(0)}, \tag{4.22}$$

$\tilde\varphi$ and φ the densities of $\widetilde Z$ and Z. See Remark 11 for the proof.

Table 6 applies Theorem 5 to the case where $\widetilde Z$ in (3.2) has a student-t distribution with 10 degrees of freedom,

$$\widetilde Z = w(Z) = T_{10}^{-1}\Phi(Z), \tag{4.23}$$

T_{10} the t_{10} cdf. Equivalence levels $\tilde\alpha$ are shown for six choices of α and nine choices of (z_0, a_0), covering a substantial range of possibilities. The match of $\tilde\alpha$ to nominal α is generally good. For $(z_0 = -0.15, a_0 = -0.1)$ we see some overcoverage at $\alpha = 0.025$ and undercoverage at $\alpha = 0.975$, with the opposite being true at $(z_0 = 0.15, a_0 = 0.1)$. All told, the bca limits look reasonably robust under GSTF model (3.2)–(4.21) but, as discussed in Remark 9, things are worse if $\widetilde Z$ in (3.2) has an asymmetric distribution.

The diagnostic functions are shown in Figure 6 for the extreme cases $(z_0, a_0) = (-0.15, -0.1)$ and $(0.15, 0.1)$ of Table 6 (along with $(0, 0)$). $D_0(z)$ is strongly nonlinear in both cases, with some effects on the accuracy of $\hat\theta_{\rm bca}[\alpha]$, but not disastrous ones.

TABLE 6
Equivalent level $\tilde{\alpha}$ corresponding to nominal NSTF level α for various choices of z_0 and a_0; student-t translation model (4.21), 10 degrees of freedom.

		Nominal level					
z_0	a_0	.025	.05	.1	.9	.95	.975
.00	.0	.025	.050	.100	.900	.950	.975
.00	.1	.026	.051	.101	.901	.951	.978
.00	−.1	.022	.049	.099	.899	.949	.974
.15	.0	.028	.053	.103	.905	.955	.979
.15	.1	.029	.054	.104	.906	.956	.984
.15	−.1	.027	.053	.103	.905	.954	.978
−.15	.0	.021	.045	.095	.897	.947	.972
−.15	.1	.022	.046	.095	.897	.947	.973
−.15	−.1	.016	.044	.094	.896	.946	.971

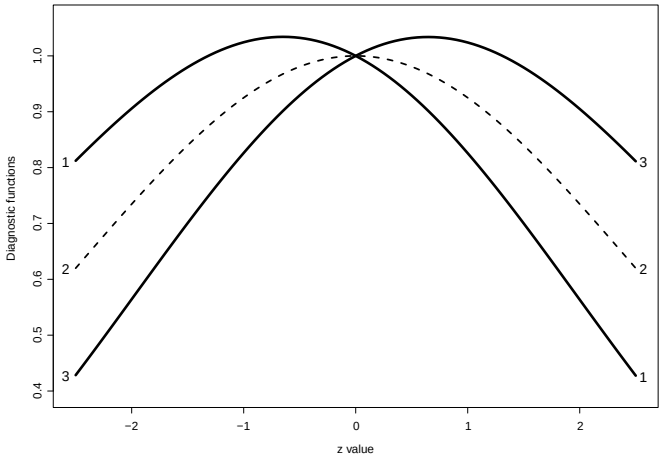

FIG 6. *Diagnostic functions for student-t translation model (4.21): (1) $z_0 = -0.15$, $a_0 = -0.10$; (2) $z_0 = 0$, $a_0 = 0$; (3) $z_0 = 0.15$, $a_0 = 0.10$.*

5. MULTIPARAMETER MODELS AND NUISANCE PARAMETERS

The discussion so far has focused on one-parameter models $F_\theta(\hat{\theta})$. However, the real need for approximate confidence intervals arises in multiparametric situations where the observed data depends on nuisance parameters as well as the parameter of interest θ. The bca confidence limits are designed to handle multidimensional problems, and this section extends the diagnostic equivalence methodology to such situations.

So far our only multiparameter calculations concerned the "maxeig" example (1.12) for the student score data. Table 3 in Section 1 showed the 95% bca interval for maxeig extending much farther to the right than the standard interval, and less far to the left. Can we believe the bca results? Figure 7 graphs the diagnostic function. It isn't perfectly linear, but the equivalence calculations in Table 7 show excellent agreement between $\tilde{\alpha}$ and α–a strong argument in favor of the bca maxeig interval, or at least for the underlying NSTF assumptions.

TABLE 7
Nominal level α and equivalent $\tilde{\alpha}$, student score maximum eigenvalue.

α	.0250	.0500	.1000	.900	.950	.975
$\tilde{\alpha}$	.0244	.0495	.0996	.899	.948	.979

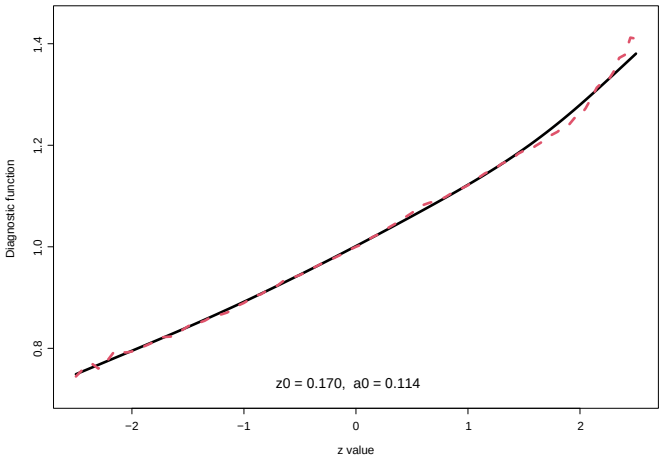

FIG 7. *Diagnostic function for maxeig, student score data; raw (dashed) and smoothed (solid).*

Figure 7 is based on an extension of the previous theory to multiparameter exponential families, those having density functions

$$g_\eta(y) = e^{\eta^\top y - \psi(\eta)} g_0(y). \tag{5.1}$$

Here η is the natural parameter and y the sufficient statistic, both η and y being p-dimensional vectors; $g_0(y)$ is the "base density", while $\exp\{\psi(\eta)\}$ is the constant that makes $g_\eta(y)$ integrate to 1. The expectation vector

$$\mu = E_\eta\{y\} \tag{5.2}$$

can be obtained as the gradient of $\psi(\eta)$; see for instance Efron (2023).

We wish to set confidence intervals for a real-valued parameter θ,

$$\theta = t(\mu), \tag{5.3}$$

for which we have the maximum likelihood point estimate

$$\hat{\theta} = t(y). \tag{5.4}$$

The normal sampling model used in Section 1,

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}_5(\gamma, \Sigma), \quad i = 1, \ldots, n = 22, \tag{5.5}$$

is of exponential family form (5.1). In this case the sufficient vector y has $p = 20$ components,

$$y = \big(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_5, \bar{x}_1^2, \bar{x}_2^2, \ldots, \bar{x}_5^2, \overline{x_1 x_2}, \overline{x_1 x_3}, \ldots, \overline{x_4 x_5}\big), \tag{5.6}$$

the bars indicating averages over the columns of the data matrix in Table 1. In Table 2, θ was the population correlation coefficient

$$\theta = \frac{\Sigma_{12}}{(\Sigma_{11}\Sigma_{22})^{1/2}}, \tag{5.7}$$

while θ was the maximum eigenvalue of Σ in Table 3.

The bca confidence algorithm used in what follows depends on *parametric bootstrap* sampling: if $\hat{\zeta}$ is the MLE of η in model (5.1), we draw B resamples of y from $g_{\hat{\zeta}}$,

$$y_1^*, y_2^*, \ldots, y_B^* \overset{\text{iid}}{\sim} g_{\hat{\zeta}}(\cdot), \tag{5.8}$$

for each one evaluating the bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}_i^* = t(y_i^*), \quad i = 1, \ldots, B. \tag{5.9}$$

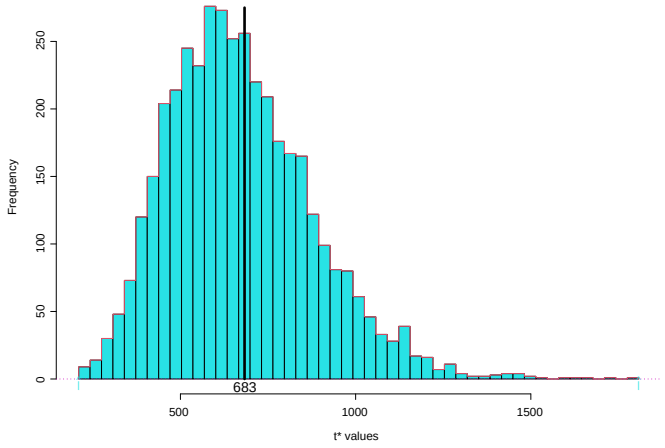

FIG 8. $B = 4000$ *normal theory parametric bootstrap replications, maxeig statistic, student score data (MLE 683).*

Figure 8 shows the histogram of $B = 4000$ resamples (at least twice more than necessary) of the maxeig statistic, obtained by sampling from model (5.5) with γ and Σ set equal to their usual maximum likelihood estimates. It is long-tailed to the right, accounting for some of the difference between the standard and bca intervals in Table 3.

Family $g_\eta(y)$ (5.1) induces a distribution for $\hat{\theta} = t(y)$, but one that usually depends on the full natural parameter η or equivalently the expectation vector μ, not just $\theta = t(\mu)$. The key step in the bca algorithm is the reduction of (5.1) to a data-based one-parameter family of distributions for $\hat{\theta}^* = t(y^*)$,

$$f_\lambda\big(\hat{\theta}_i^*\big) = e^{\lambda \dot{t}_0^\top (y_i^* - \bar{y}) - \phi(\lambda)} f_0\big(\hat{\theta}_i^*\big), \tag{5.10}$$

where $\bar{y} = \sum_{i=1}^B y_i/B$), defined for a range of the real-valued parameter λ that includes 0. Here $\bar{y}$ is considered fixed at its observed value, only y_i^* being variable.

In (5.10), $\dot{t}_0$ is the gradient vector of $t(y^*)$ evaluated at $\bar{y}$,

$$\dot{t}_0 = \left(\frac{\partial t(y^*)}{\partial y_1^*}, \frac{\partial t(y^*)}{\partial y_2^*}, \ldots, \frac{\partial t(y^*)}{\partial y_p^*}\right)\Bigg|_{y^*=\bar{y}}$$
$$\text{and} \quad \phi(\lambda) = \log\left(\sum_{i=1}^B e^{\lambda \dot{t}_0^\top (y_i^* - \bar{y})}/B\right). \tag{5.11}$$

The base density $f_0(\hat{\theta}^*)$ is discrete, putting probability B^{-1} on each value $\hat{\theta}_i^*$, $i = 1, \ldots, B$, which is to say it represents the bootstrap distribution of $\hat{\theta}^*$, as pictured for example in Figure 8.

Family (5.10) "tilts" the bootstrap density $f_0(\hat{\theta}^*)$ proportionally to $\exp\{\lambda d_i^*\}$, where

$$d_i^* = \dot{t}_0^\top (y_i^* - \bar{y}); \tag{5.12}$$

$\hat{\theta} + d_i^*$ is the first-order Taylor series approximation to $\hat{\theta}^*(y^*)$, in the direction $\dot{t}_0$. The choice of $\dot{t}_0$ as the crucial tilting direction in (5.10) is based on the *least favorable family* construction of Stein (1956); see Remark 12.

The reduction to model (5.10) puts us back in a one-parameter framework, with cdf at value t equaling

$$F_\lambda(t) = \sum_{\hat{\theta}_i^* \le t} f_\lambda\big(\hat{\theta}_i^*\big). \tag{5.13}$$

LEMMA. *The partial derivative of $F_\lambda(t)$ with respect to λ at $\lambda = 0$ is*

$$\dot{F}_0(t) = \frac{1}{B} \sum_{\hat{\theta}_i^* \le t} d_i. \tag{5.14}$$

PROOF. The derivative of $f_\lambda(\hat{\theta}_i^*) = \exp\{\lambda d_i - \phi(\lambda)\}\, B^{-1}$ with respect to λ is

$$\dot{f}_\lambda\big(\hat{\theta}_i^*\big) = (d_i - \bar{d}_\lambda) F_\lambda\big(\hat{\theta}_i^*\big), \tag{5.15}$$

where

$$\bar{d}_\lambda = \frac{\sum_1^B d_i e^{\lambda d_i}}{\sum_1^B e^{\lambda d_i}};$$

then

$$\dot{F}_\lambda(t) = \sum_{\hat{\theta}_i^* \le t} \dot{f}_\lambda\big(\hat{\theta}_i^*\big).$$

At $\lambda = 0$, $\bar{d}_\lambda = 0$ and $f_\lambda(\hat{\theta}_i^*) = B^{-1}$, giving (5.14). □

Starting from (5.14), we can compute the diagnostic function $D_0(z)$ as in (3.6)–(3.9). In Figure 7, $D_0(z)$ for the student score maxeig parameter was calculated in this way. The bca algorithm provided $z_0 = 0.170$, $a_0 = 0.114$, $\epsilon_0 = 0.116$ (3.11), $w(z)$ (4.14), and finally the equivalence values $\tilde{\alpha}$ in Table 7 (4.18). All of these calculations are based on the same bootstrap replications used

for the bca confidence limits and require almost no additional computer time.

The discussion so far concerned parametric models like (5.5), but the diagnostic theory applies just as well to nonparametric situations. Our observed data is an $n \times m$ matrix $\boldsymbol{X}$, such as the student scores in Table 1 where $n = 22$ and $m = 5$. The rows x_i are independent and identically distributed draws from some unknown distribution G on $\mathcal{R}^m$, but G can be anything at all.

The target parameter is some function $\theta = T(G)$ of the distribution G, which we estimate by $\hat{\theta} = t(\boldsymbol{X})$; very often $t(\boldsymbol{X})$ is

$$\hat{\theta} = t(\boldsymbol{X}) = T\big(\widehat{G}\big), \tag{5.16}$$

where $\widehat{G}$ is the empirical distribution

$$\widehat{G}: \text{ probability } n^{-1} \text{ on } x_i \text{ for } i = 1, \ldots, n. \tag{5.17}$$

For the correlation example (1.3), $T(\widehat{G})$ is the empirical correlation between the first two columns of $\boldsymbol{X}$, i.e., the usual Pearson correlation estimate. This is $\hat{\theta} = 0.498$, the same as the normal theory estimate (1.4), but the nonparametric and parametric confidence intervals for θ are different.

A nonparametric bootstrap data matrix $\boldsymbol{X}^*$ has its n rows chosen randomly and with replacement from the rows of $\boldsymbol{X}$. That is, its ith row x_i^* is

$$x_i^* = x_{I(j)}, \tag{5.18}$$

where $I(1), I(2), \ldots, I(n)$ is a sample of size n drawn independently with replacement from the uniform distribution on $\{1, \ldots, n\}$; $\boldsymbol{X}^*$ gives a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = t(\boldsymbol{X}^*). \tag{5.19}$$

Nonparametric bootstrap resampling falls into the multidimensional exponential family framework (5.1). The connection is through the multinomial distribution, which is an n-dimensional exponential family, with sufficient vector y equaling $(y(1), \ldots, y(n))$ where $y(j)$ is the number of times x_j occurs in the bootstrap sample. Section 5.7 of Efron (2023) provides a full discussion.

Given the data $\boldsymbol{X}$, there are a great many parametric models available but only a single nonparametric one: (5.18)–(5.19). This makes possible a single program that handles all nonparametric bca applications in a fully automatic way. Figure 9 shows diagnostic function $D_0(z)$ for the student score correlation coefficient, calculated from $B = 4000$ nonparametric bootstrap replications. It resembles the parametric function in Figure 5 of Section 3, but now with student-t translation degrees of freedom about 40 rather than 63.

Figure 10 plots the 4000 points (d_i^*, t_i^*) used for the computation in the lemma (5.14). For $z = 1$, $\alpha = 0.841$,

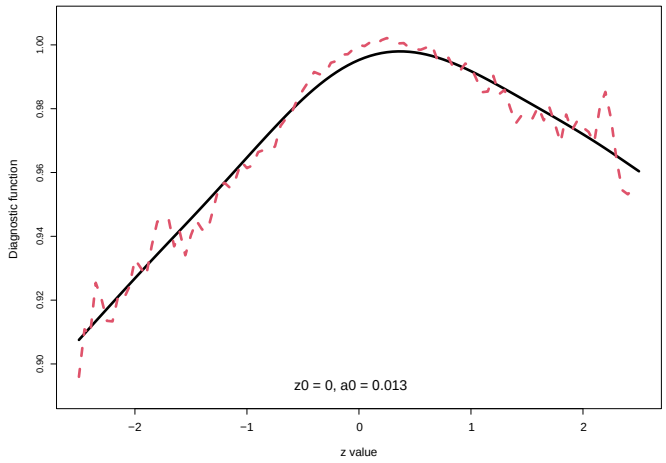

FIG 9. *Nonparametric diagnostic function for student score correlation; $B = 4000$ bootstraps; raw (dashed) and smoothed (solid).*

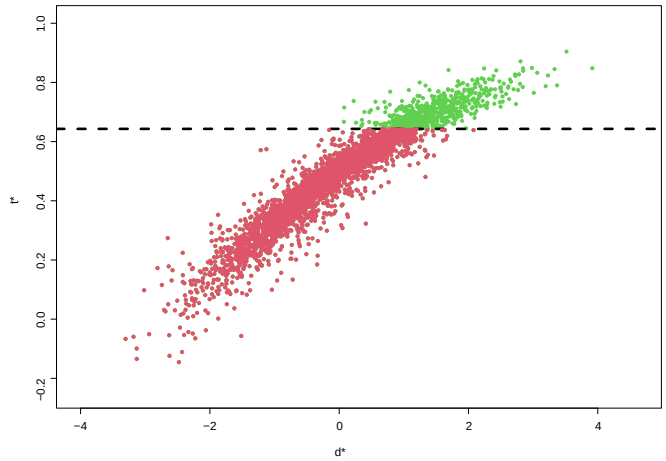

FIG 10. *t^* versus d^* for the 4000 bootstrap replications; dashed line for $D(z = 1)$.*

the bootstrap percentile $\hat{\theta}^{*(\alpha)}$ was 0.676, the height of the dashed line, illustrating the estimation of $D_0(1)$; there is almost perfect equivalence between $\tilde{\alpha}$ and α. All of this required only a single algorithmic call; see Remark 13.

Logistic regression is perhaps the most widely used parametric estimation model. Table 8 summarizes an example: 130 very sick babies at an African care facility had six health predictors measured at entry; 68 of the 130 died (see Remark 13). The investigators wished to predict which children were most at risk. Table 8 shows the output of a standard logistic regression program where the respiratory distress measure "resp" is seen to be the most predictive, with MLE point estimate $\hat{\theta} = 1.10$.

A parametric bca analysis gave these 95% confidence intervals for resp:

$$\text{bca}(0.32, 1.70) \quad \text{standard}(0.38, 1.82). \tag{5.20}$$

The bca interval was based on $B = 4000$ "Bernoulli" replications; see Remark 14. These also gave the diag-

TABLE 8
Estimates and significance levels for six predictors of death: gestational age, body weight index, respiratory rate, cpap airway blockage, mental acuity, heart rate. Logistic regression, study of 130 very sick babies at an African birth facility.

	Estimate	z-value	p-value
Gest	−.86	−1.92	.05
Bwei	.03	.08	.93
Resp	1.10	3.42	.00
CPAP	.04	.13	.89
Ment	.39	1.55	.12
Rate	−.66	−2.36	.02

TABLE 9
Equivalent GSTF levels $\tilde{\alpha}$ corresponding to nominal NSTF level α; logistic regression study.

α	.0250	.050	.100	.900	.950	.975
$\tilde{\alpha}$	.0192	.048	.098	.899	.949	.974

nostic function in Figure 11, from (5.14), and the equivalence values in Table 9. The equivalences are good but not perfect, with a tendency to overcoverage (i.e., $\tilde{\alpha} < \alpha$) for small α.

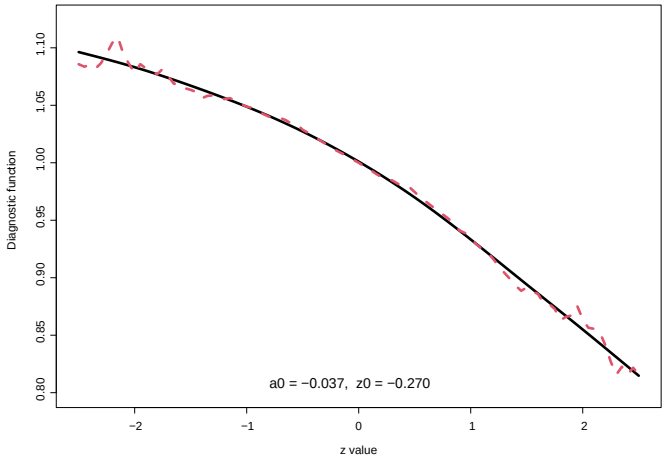

FIG 11. *Diagnostic function for logistic regression study; from $B =$ 4000 Bernoulli resamples.*

The bias correction value z_0 has almost no effect on $D_0(z)$ or $w(z)$, (4.13)–(4.14), but it can affect the bca endpoints $\hat{\theta}_{\mathrm{bca}}[\alpha]$. That is the case here, where the large negative value $z_0 = -0.270$ is reflected in the leftward shift of the bca interval.

6. REMARKS AND DISCUSSION

The remarks in this section expand on points raised in the text. A brief discussion ends the paper.

6.1 Remarks

REMARK 1 (More General GSTF Form). The original GSTF definition in Efron (1982) was broader than (3.2):

$$\hat{\phi} = \nu_\phi + \sigma_\phi w(Z) \quad (Z \sim \mathcal{N}(0,1)), \tag{6.1}$$

when ν_ϕ and $\sigma_\phi > 0$ are differentiable functions of ϕ. Our formulation (3.2) follows (6.1), with

$$\nu_\phi = \phi(1 - a_0\tilde{z}_0) \text{ and } \sigma_\phi = 1 + a_0\phi, \tag{6.2}$$

or, letting $\tilde{\phi} = \phi(1 - a_0\tilde{z}_0)$,

$$\nu_{\tilde{\phi}} = \tilde{\phi} \text{ and } \sigma_{\tilde{\phi}} = 1 + \epsilon_0\tilde{\phi} \tag{6.3}$$

where $\epsilon_0 = a_0(1 - a_0\tilde{z}_0)^{-1}$ as in (3.11).

REMARK 2 (Converse to the Corollary to Theorem 2). The corollary states that an NSTF (2.1)–(2.2) has linear diagnostic function $D_\theta(z) = 1 + z\epsilon_0$. Conversely, if $D_\theta(z)$ is linear with slope ϵ_0, then the only GSTF having that value of ϵ_0 is an NSTF. This is immediate from Theorem 4: $D_\theta(z) = 1 + z\epsilon_0$ given

$$I(z) = \frac{1}{\epsilon_0}\log(1 + \epsilon_0 z) \tag{6.4}$$

from (4.12), and $w(z) = z$ from (4.14); that is, $\widetilde{Z} = Z$, which is the definition of an NSTF. There *are* non-NSTF choices giving $D_\theta(z) = 1 + z\epsilon_0$, but they must have parameter ϵ_0 in (4.14) not equaling the slope ϵ_0. (See Remark 8.)

REMARK 3 (Scale and Translation Families). Suppose

$$\hat{\theta} = \theta X \quad (\theta > 0), \tag{6.5}$$

where X is a positive random variable with known cdf $F_1(x)$. For the gamma scale family (3.13)–(3.14), X was G_ν/ν. The cdf of $\hat{\theta}$ given θ is

$$F_\theta\big(\hat{\theta}\big) = F_1\big(\hat{\theta}/\theta\big), \tag{6.6}$$

so that

$$\dot{F}_\theta\big(\hat{\theta}\big)\Big|_{\theta=1} = -\hat{\theta} f_1\big(\hat{\theta}\big), \tag{6.7}$$

where $f_1(x)$ is the density of X.

Following through definitions (3.7)–(3.9), the diagnostic function is

$$D_\theta(z^{(\alpha)}) = \frac{x^{(\alpha)} f_1(x^{(\alpha)})/\varphi(z^{(\alpha)})}{x^{(0.5)} f_1(x^{(0.5)})/\varphi(0)}, \tag{6.8}$$

not depending on θ.

Figure 4 concerns a *translation model*,

$$\hat{\theta} = \theta + X, \tag{6.9}$$

with X a student-t variate in (3.20). Then $F_\theta(\hat\theta) = F_1(\hat\theta - \theta)$ leads to the diagnostic function

$$D_\theta(z^{(\alpha)}) = \frac{f_1(x^{(\alpha)})/\varphi(z^{(\alpha)})}{f_1(x^{(0.5)})/\varphi(0)}, \tag{6.10}$$

graphed in Figure 4.

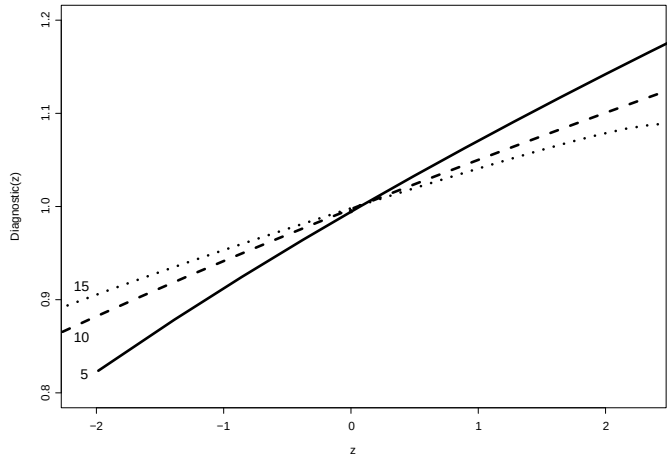

FIG 12. *Poisson distribution diagnostic functions for $\mu = 5, 10$, and 15.*

REMARK 4 (One-parameter Exponential Families). A one-parameter exponential family $X \sim f_\theta(x)$ has densities

$$f_\theta(x) = e^{\theta x - \psi(\theta)} f_0(x), \quad x \in (x_{\text{lo}}, x_{\text{up}}), \tag{6.11}$$

where the range $(x_{\text{lo}}, x_{\text{up}})$ may be infinite. (The gamma scale family (3.5) can be rewritten in form (6.11).) The derivative of $f_\theta(x)$ with respect to θ is

$$\dot f_\theta(x) = (x - \mu_\theta) f_\theta(x), \tag{6.12}$$

where $\mu_\theta = \dot\psi(\theta)$ is the expectation parameter $E_\theta\{X\}$. This leads to the following expression for the diagnostic function:

$$D_\theta(z^{(\alpha)}) = \frac{\int_{x_{\text{lo}}}^{x^{(\alpha)}} (x - \mu_\theta) f_\theta(x)\, dx}{\int_{x_{\text{lo}}}^{x^{(0.5)}} (x - \mu_\theta) f_\theta(x)\, dx}, \tag{6.13}$$

closely related to the lemma of Section 5 (5.14).

The Poisson family,

$$f_\mu(x) = e^{-\mu}\mu^x/x! \text{ for } x = 0, 1, \ldots, \tag{6.14}$$

can be written in form (6.11). Figure 12 shows diagnostic function $D_\mu(z)$ for $\mu = 5, 10$, and 15. Here the discreteness of (6.14) has been taken into account by a "half-atom" correction to the cdf,

$$F_\mu(x) = \sum_{y=0}^{x-1} f_\mu(y) + \frac{1}{2} f_\mu(x). \tag{6.15}$$

TABLE 10
Weibull scale parameter family; nominal NSTF levels α and actual equivalent GSTF levels $\tilde\alpha$.

α	.0250	.0500	.100	.900	.950	.975
$\tilde\alpha$	.0254	.0503	.100	.941	.941	.941

$D_\mu(z)$ is close to linear even for $\mu = 5$, where definition (6.15) is most questionable.

In one-parameter exponential families, including the Poisson, both z_0 and a_0, (2.7) and (2.9), are closely approximated by

$$z_0 = a_0 = \text{skewness}(x)/6. \tag{6.16}$$

These determine ϵ_0, which together with $D_\mu(z)$ provides the equivalence table. For $\mu = 10$ and $\mu = 15$, $\tilde\alpha$ matched α to three decimal places; for $\mu = 5$, $\tilde\alpha = 0.024$ for $\alpha = 0.025$, with similar errors at $\alpha = 0.05, 0.10, 0.9, 0.95$ and 0.975. Exact intervals are available for the Poisson, for which the bca limits give about two-place accuracy.

REMARK 5 (Weibull Scale Family). A scaled Weibull variate $\hat\theta$ takes the form

$$\hat\theta = \theta X^{1/c}, \tag{6.17}$$

where c has a known positive value and X has a Gamma distribution with one degree of freedom, $f(x) = \exp(-x)$ for $x > 0$. The monotone mappings $\hat\gamma = \hat\theta^c$ and $\gamma = \theta^c$ transform (6.17) to

$$\hat\gamma = \gamma X, \tag{6.18}$$

a Gamma scale family with degrees of freedom equal 1, where, from (6.16),

$$z_0 = a_0 = 0.337. \tag{6.19}$$

The equivalence table in Table 10 looks fine on the left but disastrous on the right. $D_\theta(z)$ is highly curved in this case — roughly six times more curved than the $\nu = 5$ line in Figure 3 — but more of the trouble comes from the large values of z_0 and a_0. The "corrected z-value"

$$Z = \frac{z_0 + (z_0 + z^{(\alpha)})}{1 - a_0(z_0 + z^{(\alpha)})} \tag{6.20}$$

in the bca formula (2.3) is 64.9 for $\alpha = 0.975$, far beyond the reasonable range, say $|Z| < 4$, for the bca corrections to the standard intervals. A confidence interval is an ambitious enterprise in that it attempts to account for both bias and variability; as mentioned earlier, the bca formula can be pushed too far, particularly if the bias corrections need to be enormous.

REMARK 6 (A Case Where NSTF Equals GSTF). Suppose $a_0 = 0$, $p_0 = F_{\hat\theta}(\hat\theta) = 0.50$, and $\widetilde Z$ is symmetrically distributed around zero (as it is for the student-t model (4.21)). Then $\tilde\alpha = \alpha$, i.e., $\hat\theta_{\text{bca}}[\alpha]$ is the same as the GSTF confidence limit $\hat\theta_{\text{gbca}}[\alpha]$.

PROOF. We have $a_0 = 0$ and $\tilde{z}_0 = \widetilde{\Phi}^{-1}(p_0) = 0$, so (4.4) gives

$$\tilde{\beta}[\alpha] = \widetilde{\Phi}\big\{-\widetilde{\Phi}(1-\alpha)\big\} = \widetilde{\Phi}\big\{\widetilde{\Phi}^{-1}(\alpha)\big\} = \alpha, \tag{6.21}$$

which implies $\tilde{\beta}[\alpha] = \beta[\alpha] = \alpha$. Here we have used $-\widetilde{\Phi}(1-\alpha) = \widetilde{\Phi}(\alpha)$ by symmetry. □

The top row of Table 6 illustrates this result.

REMARK 7 (Proof of Theorem 3). Working on the transformed scale ϕ (2.1)–(2.2), we wish to show that the generalized bca confidence limit $\hat{\phi}_{\mathrm{gbca}}[\alpha]$(4.5) is accurate and correct given the GSTF model (3.2); that is, that

$$\hat{\phi}_{\mathrm{gbca}}[\alpha] = F_{\hat{\phi}}^{-1}\big(\tilde{\beta}[\alpha]\big) \tag{6.22}$$

is the appropriate confidence level limit, where $F_{\hat{\phi}}$ is the cdf of $\hat{\phi}^*$ if $\phi = \hat{\phi}$, and

$$\tilde{\beta}[\alpha] = \widetilde{\Phi}\left\{\tilde{z}_0 + \frac{\tilde{z}_0 - \tilde{z}^{(1-\alpha)}}{1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})}\right\}. \tag{6.23}$$

The proof depends on the transformation invariance of confidence intervals, bootstrap distributions, and bca formulas under monotone mappings.

The GSTF model $\hat{\phi} = \phi + (1 + a_0\phi)(\widetilde{Z} - \tilde{z}_0)$ can be written as

$$\begin{aligned} \hat{\gamma} &= \gamma R \quad \text{where} \\ \hat{\gamma} &= 1 + a_0\hat{\phi}, \\ \gamma &= 1 + a_0\phi, \quad \text{and} \\ R &= 1 + a_0(\widetilde{Z} - \tilde{z}_0), \end{aligned} \tag{6.24}$$

an exponentiated form of (4.7). This is a scale parameter model: having observed $\hat{\phi}$, the accurate and correct level α confidence limit for γ is

$$\hat{\gamma}[\alpha] = \hat{\gamma}/R^{(1-\alpha)}. \tag{6.25}$$

Define β as the value satisfying

$$R^{(\beta)} = 1/R^{(1-\alpha)}, \tag{6.26}$$

so $\hat{\gamma}[\alpha] = \hat{\gamma}R^{(\beta)}$ for any choice of β.

The parametric bootstrap distribution of $\hat{\gamma}^*$ given $\hat{\gamma}$ is

$$\hat{\gamma}^* \sim \hat{\gamma}R, \tag{6.27}$$

so

$$\hat{\gamma}[\alpha] = \hat{\gamma}R^{(\beta)} = \hat{\gamma}^{*(\beta)}, \tag{6.28}$$

the βth bootstrap percentile of $\hat{\gamma}^*$. Transformation invariance implies that

$$\hat{\gamma}[\alpha] = \hat{\phi}^{*(\beta)} \tag{6.29}$$

for any value of $\hat{\phi}$. Since $\hat{\phi}^{*(\beta)} = F_{\hat{\phi}}^{-1}(\beta)$ (given an infinite number of bootstrap replications), the proof of Theorem 3 requires showing that $\tilde{\beta}[\alpha]$ (6.23) equals β.

TABLE 11
Equivalence table if $D(z) = 1 + 0.1z$ but true value of ϵ is zero. NSTF intervals are too long on the left and too short on the right.

α	.0250	.0500	.1000	.900	.950	.975
$\tilde{\alpha}$	.0074	.0245	.0708	.872	.921	.949

Consider the case $\hat{\phi} = 0$, $\hat{\gamma} = 1$, for which $R^{(1-\alpha)} = 1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})$,

$$\hat{\gamma}[\alpha] = \Big[1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})\Big]^{-1}, \tag{6.30}$$

and

$$\hat{\phi}[\alpha] = \frac{1}{a_0}\big(\hat{\gamma}[\alpha] - 1\big) = \frac{\tilde{z}_0 - \tilde{z}^{(1-\alpha)}}{1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})}. \tag{6.31}$$

But for $\hat{\phi} = 0$, $\hat{\phi}^* \sim \widetilde{Z} - \tilde{z}_0$ has

$$\hat{\phi}^{*(\beta)} = \widetilde{Z}^{(\beta)} - \tilde{z}_0 = \widetilde{\Phi}^{-1}(\beta) - \tilde{z}_0 \tag{6.32}$$

(remembering that $\widetilde{\Phi}$ is the cdf of $\widetilde{Z}$). Combining (6.31) and (6.32) gives

$$\widetilde{\Phi}^{-1}(\beta) = \tilde{z}_0 + \frac{\tilde{z}_0 - \tilde{z}^{(1-\alpha)}}{1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})} \tag{6.33}$$

or

$$\beta = \widetilde{\Phi}\left\{\tilde{z}_0 + \frac{\tilde{z}_0 - \tilde{z}^{(1-\alpha)}}{1 - a_0(\tilde{z}_0 - \tilde{z}^{(1-\alpha)})}\right\} = \tilde{\beta}[\alpha], \tag{6.34}$$

showing that $\hat{\phi}_{\mathrm{gbca}}[\alpha] = F_{\hat{\phi}}^{-1}(\tilde{\beta}[\alpha])$. By transformation invariance, this implies $\hat{\theta}_{\mathrm{gbca}}[\alpha] = \widehat{F}^{-1}(\tilde{\beta}[\alpha])$, verifying Theorem 3.

REMARK 8 ($D_\theta(z)$ and Unidentifiability). According to Theorem 4 (4.13)–(4.14), $D_\theta(z)$ (which determines $I(z)$) and ϵ_0 together determine the function $w(z)$, which in turn determines the equivalence calculations. $D_\theta(z)$ by itself, however, does *not* provide $w(z)$: for any possible value of ϵ_0, $w(z)$ as given by (4.13)–(4.14) produces the same $D_\theta(z)$ (3.10).

Suppose, for instance, $D_\theta(z) = 1 + 0.1 \cdot z$, the diagnostic function for an NSTF having $\epsilon_0 = a_0(1 - a_0 z_0)^{-1} = 0.1$. However, $D_\theta(z)$ is also the diagnostic function for a GSTF having ϵ_0 value zero and

$$w(z) = I(z) = \frac{1}{0.1}\log(1 + 0.1 \cdot z), \tag{6.35}$$

as in (6.4) and (4.13). If in fact the GSTF model (6.35) was correct, the NSTF confidence limits $\hat{\theta}_{\mathrm{bca}}[\alpha]$ based on $\epsilon_0 = 0.1$ would be terrible, as shown by Table 11. The bca algorithm provides estimates of an appropriate value of ϵ_0, as well as $D_\theta(z)$, helping avoid this kind of problem.

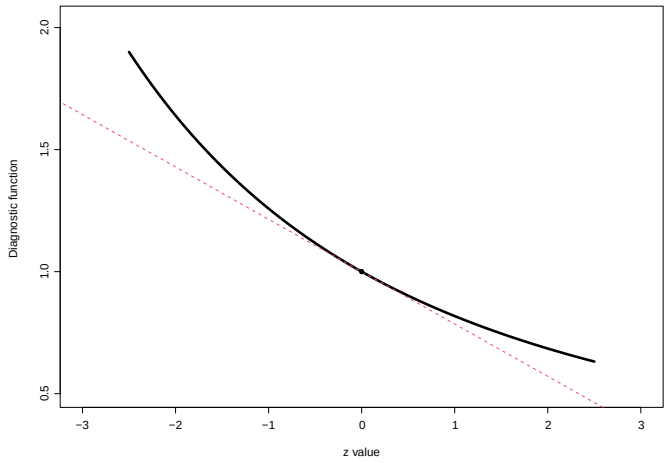

FIG 13. *Diagnostic function, Gamma translate* $\nu = 10$ *(dashed is tangent line).*

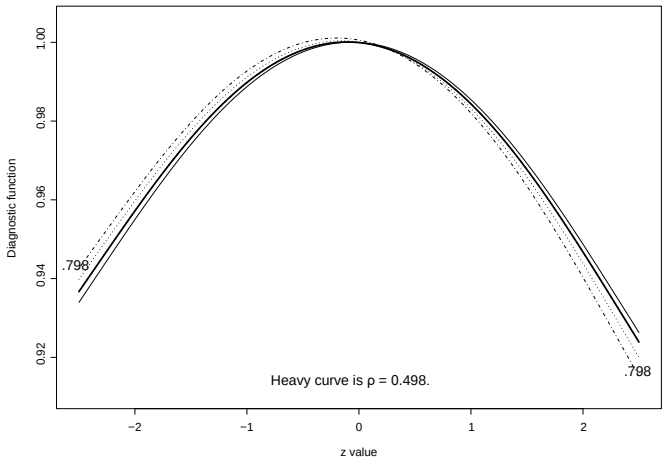

FIG 14. *Diagnostic function normal correlation coef* $n = 22$*; for observed* $\rho = 0.298, 0.498, 0.698, 0.798$.

TABLE 12
Equivalence table for gamma translate model (6.34)*. NSFT intervals are much too short on the left and long on the right.*

α	.025	.050	.100	.900	.950	.975
$\tilde{\alpha}$	.086	.114	.158	.963	.994	.994

REMARK 9 (An Asymmetric GSTF). Suppose we have a gamma translation model,

$$\hat{\theta} = \theta + Q \quad (Q \sim G_\nu), \tag{6.36}$$

G_ν gamma with ν degrees of freedom. (*Not* the gamma scale model (3.13).) This is already in the GSTF form (4.7), with $\theta = \zeta$, $\hat{\theta} = \hat{\zeta}$, $z_0 = 0$, and $a_0 = 0$. The GSTF limits are exact in this case,

$$\hat{\theta}_{\mathrm{gbca}}[\alpha] = \hat{\theta} - Q^{(1-\alpha)}, \tag{6.37}$$

but the NSTF limits are different,

$$\hat{\theta}_{\mathrm{bca}}[\alpha] = \hat{\theta} + Q^{(\alpha)}. \tag{6.38}$$

The asymmetry, $G_\nu^{(\alpha)} \neq -G_\nu^{(1-\alpha)}$, is severe for $\nu = 10$. Model (6.36) gives the curved diagnostic function seen in Figure 13. Carrying through the equivalence calculations shows how badly $\hat{\theta}_{\mathrm{bca}}[\alpha]$ performs in this case, as shown in Table 12. In a genuine data analysis, Table 12 would fulfill its purpose by warning against the bca intervals.

REMARK 10 ($D_\theta(z)$ Not Depending on θ). The fact that a GSTF can be transformed into a translation family (4.7) implies that its diagnostic functions $D_\theta(z)$ do not depend on θ. Figure 14 shows $D_\theta(z)$ for θ equaling 0.298, 0.498, 0.698 and 0.798, for a normal correlation coefficient with $n = 22$. There *is* a small amount of dependence on θ (though not enough to affect the equivalence tables), showing that $F_\theta(\hat{\theta})$ isn't exactly GSTF.

REMARK 11 (Proof of Theorem 5). Family (4.19) has cdfs $F_\phi(\hat{\phi}) = \widetilde{\Phi}[(\hat{\phi} - \phi)\sigma_\phi^{-1} + z_0]$. Differentiating with respect to ϕ gives

$$\dot{F}_\phi(\hat{\phi}) = \frac{-(1 + a_0\hat{\phi})}{\sigma_\phi^2}\tilde{\varphi}\left(\frac{\hat{\phi} - \phi}{\sigma_\phi} + \tilde{z}_0\right). \tag{6.39}$$

There is no loss of generality in taking $\phi = 0$, $\sigma_\phi = 1$. (If actually $\phi = \phi_0$, redefining the transformation $m(\theta)$ from $\phi = m(\theta)$ to $(m(\theta) - \phi_0)\sigma_{\phi_0}^{-1}$ makes $\phi = 0$.) Then (6.39) becomes

$$\dot{F}_0(\hat{\phi}) = -(1 + a_0\hat{\phi})\tilde{\varphi}(\hat{\phi} + \tilde{z}_0). \tag{6.40}$$

For $\phi = 0$ the α percentile $\hat{\phi}^{(\alpha)} = \widetilde{Z}^{(\alpha)} - \tilde{z}_0$, giving

$$\frac{\dot{F}_0(\hat{\phi}^{(\alpha)})}{\varphi(z^{(\alpha)})} = \frac{-\left[1 + a_0(\widetilde{Z}^{(\alpha)} - \tilde{z}_0)\right]\tilde{\varphi}(\widetilde{Z}^{(\alpha)})}{\varphi(z^{(\alpha)})}, \tag{6.41}$$

and, from (3.7)–(3.9),

$$D_0(z) = \frac{1 + a_0(\widetilde{Z}^{(\alpha)} - \tilde{z}_0)}{1 - a_0\tilde{z}_0}r(z), \tag{6.42}$$

which is the same as (4.21)–(4.22).

REMARK 12 (Stein's Least Favorable Family). For points y_i^* near $\bar{y}$, $\theta_i^* = t(y_i^*)$ changes most rapidly in direction $\dot{t}_0$ (5.11). Stein (1956) showed that in family (5.10) the variance of $\hat{\theta}^*$ at $\lambda = 0$ is *not less* than the Fisher information bound for estimating θ in the full p-dimensional exponential family. Any choice of tilting direction other than $\dot{t}_0$ makes the one-dimensional variance smaller, which is to say that $\dot{t}_0$ is least favorable. Section 6 of Efron (1987) discusses the least favorable family's role in the bca algorithm.

TABLE 13

Partial output of nonparametric bootstrap confidence interval program `bcadiag`, $B = 4000$, *applied to respiratory predictor of sick babies study, Table 8. First two columns are an equivalence table that shows close correspondence between nominal NSTF level* α *and GSFT level* $\tilde{\alpha}$, *except at* $\alpha = 0.025$ *where NSTF endpoint is conservative. Last two columns compare NSTF and standard endpoints. Column* β *shows percentiles of 4000 bootstrap replications giving bcalims.*

α	$\tilde{\alpha}$	β	bcalims	Standard
.025	.014	.005	.311	.277
.050	.050	.014	.406	.409
.100	.101	.037	.536	.561
.160	.162	.070	.644	.681
.500	.500	.329	1.015	1.096
.840	.841	.704	1.399	1.512
.900	.901	.791	1.506	1.632
.950	.951	.875	1.671	1.783
.975	.975	.925	1.833	1.915

REMARK 13 (Computer Programs). The R program `bcadiag`, available from the author, performs nonparametric bca/diagnostic calculations. Table 13 shows part of the output of the call

$$\texttt{bcadiag (4000,}\boldsymbol{X}\texttt{,func)}, \tag{6.43}$$

where $B = 4000$ was the number of nonparametric bootstrap replications; $\boldsymbol{X}$ was the 130×7 data matrix for the sick babies study of Table 8, having ith row the six predictors and response (lived or died) for baby i; and `func`$(\boldsymbol{X}^*)$ computed $\hat{\theta}^*$, the "resp" coefficient estimated from bootstrap data matrix $\boldsymbol{X}^*$. (Running time was about 2.5 seconds.)

The equivalence levels $\tilde{\alpha}$ nearly match the nominal bca levels α except at $\alpha = 0.025$, where $\hat{\theta}_{\rm bca}[\alpha]$ is conservative–similarly to the parametric results in Table 9. Call (6.43) also returns the diagnostic function $D_\theta(z)$ and $w(z)$ (3.2).

Program `bcadiag` is completely automatic in the sense described in Section 1: it applies to all one-sample nonparametric situations without requiring any more from the user beyond `func`. Parametric bootstrap calculations are less automatic. They were carried out for this paper using augmented versions of algorithm `bcapar` from the CRAN package `bcaboot`; the algorithm requires preliminary calculation of bootstrap replications, as described in Efron and Narasimhan (2020).

REMARK 14 (Logistic Regression of the Sick Babies Data). The "Bernoulli" bootstrap replications used for Figure 11 and Table 9 were carried out as follows: the initial logistic regression gave estimated probability $\hat{\pi}_i$ for the ith baby's death; bootstrap response vector $\boldsymbol{y}^*$ then took

$$y_i^* \overset{\text{ind}}{\sim} \text{bern}(\hat{\pi}_i) \ \text{ for } i = 1, \ldots, 130; \tag{6.44}$$

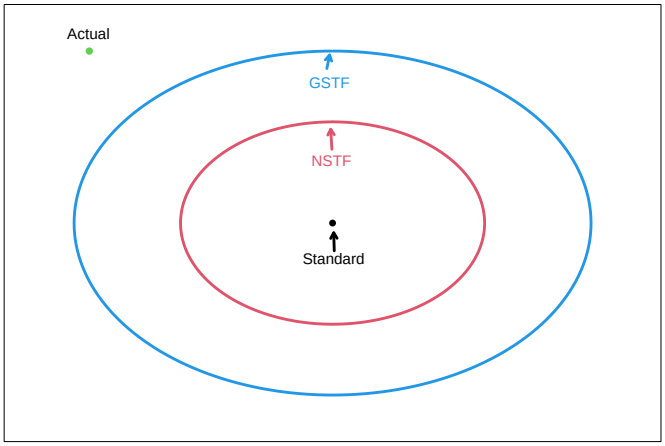

FIG 15. *Three possible target classes for setting approximate confidence intervals.*

logistic regression of $\boldsymbol{y}^*$ on $\boldsymbol{x}$, the original 130×6 matrix of predictions, gave bootstrap replications resp*. The data for Table 8 is part of a larger study reported in Mediratta et al. (2020).

6.2 Discussion

Applied statisticians depend heavily on an extended version of the central limit theorem in which estimators $\hat{\theta} = t(\boldsymbol{x})$ typically approach normality,

$$\hat{\theta} \doteq \mathcal{N}(\theta, \sigma^2), \tag{6.45}$$

as the information in $\boldsymbol{x}$ grows asymptotically large. Just what constitutes "large" depends on the situation. For the student score correlation of Section 1, $n = 22$ was not large enough, with Figure 1 making clear the nonnormality of $\hat{\theta}$. Fisher's transformation (1.9) greatly accelerates convergence to normality. The bca method automates Fisher's method, effectively reducing by an order of magnitude the sample size needed for the accuracy of the central limit theorem calculations.

Justification for the bca intervals depends on the NSTF structure (2.1)–(2.2). The paper develops data-based methods — diagnostic functions and equivalence tables — for assessing the plausibility of the NSTF assumptions. A larger class of possible models, GSTF (3.2), is used to estimate the actual coverage claimed for a given bca interval.

Figure 15 portrays the setting of approximate confidence intervals in schematic terms. Three increasingly broad classes of possible assumptions are pictured, each of which suggests its confidence interval formula. The standard intervals (1.1), by far the most widely used, depend on the normal translation class (6.45), represented by the dot at the center of Figure 15. The much larger NSTF class leads to the bca intervals (2.3). Still larger, the GSTF class gives the gbca intervals (4.5).

Presented with an actual estimation problem (the "Actual" point in Figure 15), for instance the sick babies logistic regression example in Section 5, we don't expect any of the classes to fit perfectly, but hope to be close enough to one of them to get an accurate interval. The diagnostic/equivalence calculations of this paper offer a guide to the accuracy of a given bca interval. Figure 5, for example, suggests that NSTF is a close but not perfect fit for the student score correlation problem, while equivalence Table 5 offers reassurance for using the bca formula.

In the schematic diagram, *Actual* lies closer to the NSTF oval than the Standard point. This has to be true in an asymptotic sense since bca gives second-order accurate intervals, compared to the first-order standard intervals. GSTF fills an even larger oval, but whether or not it gives still higher asymptotic accuracy isn't known; its purpose here has been as a testbed for the performance, good or bad, of bca intervals. These generally performed well in my examples, but the method can be pushed too far (as in Remark 5), in which case the diagnostic/equivalence computations will provide useful warnings.

Fisher criticized confidence intervals as giving exact but possibly incorrect inferences. If one believes model (6.45) then the standard intervals $\hat{\theta} \pm z^{(\alpha)}\sigma$ seem unassailably correct, at least in the absence of Bayesian prior information. In this same sense, $\hat{\theta}_{\mathrm{bca}}[\alpha]$ and $\hat{\theta}_{\mathrm{gbca}}[\alpha]$ are correct within the NSTF and GSTF frameworks. Accuracy concerns how fast *Actual* approaches a destination in Figure 15 (faster for NSTF than standard), while correctness concerns the appropriateness of the destination.

Davison and Hinkley (1997) describes a collection of bootstrap confidence interval methods other than bca. Beginning with an important 1983 paper by Barndorff-Nielsen (1983), methods based on higher-order expansions of the likelihood function have permitted second- and even third-order accurate intervals, going beyond bca to allowing conditioning on ancillary statistics. See Pierce and Bellio (2017). The focus here on bca reflects its relatively simple logical structure, computational ease (in the computer era), and its usual good performance in applications. "Usual" isn't always, though, and the methods of this paper allow a data-based check on coverage accuracy.

REFERENCES

BARNDORFF-NIELSEN, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365. https://doi.org/10.1093/biomet/70.2.343 MR712023

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511802843 MR1478673

DICICCIO, T. J. and EFRON, B. (1996). Bootstrap confidence intervals. *Statist. Sci.* **11** 189–228. With comments and a rejoinder by the authors. MR1436647 (98i:62042)

EFRON, B. (1982). Transformation theory: How normal is a family of distributions? *Ann. Statist.* **10** 323–339. MR653511 (84i:62018a)

EFRON, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82** 171–200. With comments and a rejoinder by the author. MR883345 (88m:62053)

EFRON, B. (2023). *Exponential Families in Theory and Practice*. *Institute of Mathematical Statistics Textbooks*. Cambridge University Press. https://doi.org/10.1017/9781108773157

EFRON, B. and NARASIMHAN, B. (2020). The automatic construction of bootstrap confidence intervals. *J. Comput. Graph. Stat.* **29** 608–619. https://doi.org/10.1080/10618600.2020.1714633

HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.* **16** 927–985. https://doi.org/10.1214/aos/1176350933 MR959185 (89h:62085)

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR560319 (81h:62003)

MEDIRATTA, R. P., AMARE, A. T., BEHL, R., EFRON, B., NARASIMHAN, B., TEKLU, A., SHEHIBO, A., AYALEW, M. and KACHE, S. (2020). Derivation and validation of a prognostic score for neonatal mortality in Ethiopia: A case-control study. *BMC Pediatr.* **20** 238. https://doi.org/10.1186/s12887-020-02107-8

PIERCE, D. A. and BELLIO, R. (2017). Modern likelihood-frequentist inference. *Int. Stat. Rev.* **85** 519–541. https://doi.org/10.1111/insr.12232 MR3723615

STEIN, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I* 197–206. University of California Press, Berkeley and Los Angeles. MR0084922 (18,948c)