The Bootstrap Small Sample Properties[1]

F.W. Scholz
University of Washington

June 25, 2007

[1]edited version of a technical report of same title and issued as bcstech-93-051, Boeing Computer Services, Research and Technology.

Abstract

This report reviews several bootstrap methods with special emphasis on small sample properties. Only those bootstrap methods are covered which promise wide applicability. The small sample properties can be investigated analytically only in parametric bootstrap applications. Thus there is a strong emphasis on the latter although the bootstrap methods can be applied nonparametrically as well. The disappointing confidence coverage behavior of several, computationally less extensive, parametric bootstrap methods should raise equal or even more concerns about the corresponding nonparametric bootstrap versions. The computationally more expensive double bootstrap methods hold great hope in the parametric case and may provide enough assurance for the nonparametric case.

Contents

1 The General Bootstrap Idea

1.1 Introduction

The bootstrap method was introduced by Efron in 1979. Since then it has evolved considerably. Efron's paper has initiated a large body of hard theoretical research (much of it of asymptotic or large sample character) and it has found wide acceptance as a data analysis tool. Part of the latter is due to its considerable intuitive appeal, which is in contrast to the often deep mathematical intricacies underlying much of statistical analysis methodology. The basic bootstrap method is easily grasped by practitioners and by consumers of statistics.

The popularity of the bootstrap was boosted early on by the very readable *Scientific American* article by Diaconis and Efron (1983). Having chosen the catchy name "bootstrap" certainly has not hurt its popularity. In Germany one calls the bootstrap method "die Münchhausen Methode," named after Baron von Münchhausen, a fictional character in many phantastic stories. In one of these he is supposed to have saved his life by pulling himself out of a swamp by his own hairs. The first reference to "die Münchhausen Methode" can be traced to the German translation of the Diaconis and Efron article, which appeared in *Spektrum der Wissenschaft* in the same year. There the translator recast the above episode to the following iamge: Pull yourself by your mathematical hairs out of the statistical swamp.

Hall (1992) on page 2 of his extensive monograph on the bootstrap expresses these contrasting thoughts concerning the "bootstrap" name:

> Somewhat unfortunately, the name "bootstrap" conveys the impression of "something for nothing" — of statisticians idly resampling from their samples, presumably having about as much success as they would if they tried to pull themselves up by their bootstraps. This perception still exists in some quarters. One of the aims of this monograph is to dispel such mistaken impressions by presenting the bootstrap as a technique with a sound and promising theoretical basis.

Much of the bootstrap's strength and acceptance also lies in its versatility. It can handle a very wide spectrum of data analysis situations with equal ease. In fact, it facilitates data anlyses that heretofore were simply impossible because the obstacles in the mathematical analysis were just too forbidding.

This gives us the freedom to model the data more accurately and obtain approximate answers to the right questions instead of right answers to often the wrong questions. This freedom is bought at the cost of massive simulations of resampled data sets followed by corresponding data analyses for each such data set. The variation of results obtained in these alternate data analyses should provide some insight into the accuracy and uncertainty of the data analysis carried out on the original data.
This approach has become feasible only because of the concurrent advances in computing. However, certain offshoots of the bootstrap, such as iterated bootstrap methods, can still strain current computing capabilities and efficient computing strategies are needed.
As stated above, the bootstrap has evolved considerably and there is no longer a single preferred method, but a wide spectrum of separate methods, all with their own strengths and weaknesses. All of these methods generally agree on the same basic bootstrap idea but differ on how they are implemented.
There are two major streams, namely the *parametric bootstrap* and the *nonparametric bootstrap*, but even they can be viewed in a unified fashion. The primary focus of this report is on parametric bootstrap methods, although the definitions for the various bootstrap methods are general enough to be applicable for the parametric and nonparametric case. The main reason for this focus is that in certain parametric examples one can examine analytically the small sample properties of the various bootstrap methods. Such an analysis is not possible for the nonparametric bootstrap.

1.2 Setup and Objective

We begin by assuming a very generic data analysis situation, namely that we have some data set $\mathbf{X}$. Data are uncertain for various reasons (sampling variability, measurement error, etc.) and we agree that the data set was generated by a probability mechanism which we denote by P. We do not know P, but through $\mathbf{X}$ we get some indirect information about P. Much of statistical inference consists in using $\mathbf{X}$ to make some inference concerning the particulars of P.
A very common structure for $\mathbf{X}$ is that it represents some random sample, i.e., $\mathbf{X} = (X_1, \ldots, X_n)$ and the X_i are independent and identically distributed (i.i.d.). Other structures involve known covariates, which can be thought of as being a known part of the specified probability model. By keeping the

data set as generic as possible we wish to emphasize the wide applicability of the bootstrap methods.
Not knowing P is usually expressed by stating that P is one of many possible probability mechanisms, i.e., we say that P is a member of a family $\mathcal{P}$ of probability models that could have generated $\mathbf{X}$.
In the course of this report we will repeatedly use specific examples for probability models and for ease of reference we will list most of them here.

Example 1: $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample of size n from some distribution function $F \in \mathcal{F}$, the family of all such distribution functions on the real line, and let $\mathcal{P} = \{P_F : \ F \in \mathcal{F}\}$. We say that $\mathbf{X} = (X_1, \ldots, X_n)$ was generated by P_F (and write $\mathbf{X} \sim P_F$) if $X_1, \ldots, X_n$ are independent, each having the same distribution function F.

Example 2: $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample of size n from a normal population with mean μ and variance σ^2. Let $\mathcal{F}$ be the family of all normal distributions, with $\mu \in R$ and $\sigma > 0$ and let $\mathcal{P} = \{P_F : \ F \in \mathcal{F}\}$. We say that $\mathbf{X} = (X_1, \ldots, X_n)$ was generated by P_F (and write $\mathbf{X} \sim P_F$) if $X_1, \ldots, X_n$ are independent, each having the same normal distribution function $F \in \mathcal{F}$.

Example 3: $\mathbf{X} = \{(t_1, Y_1), \ldots, (t_n, Y_n)\}$, where $t_1, \ldots, t_n$ are fixed known constants (not all equal) and $Y_1, \ldots, Y_n$ are independent random variables, which are normally distributed with common variance σ^2 and respective means $\mu(t_1) = \alpha + \beta t_1, \ldots, \mu(t_n) = \alpha + \beta t_n$, i.e., we write $Y_i \sim \mathcal{N}(\alpha + \beta t_i, \sigma^2)$. Here α, β, and $\sigma > 0$ are unknown parameters. Let $\mathcal{F} = \{F = (F_1, \ldots, F_n) : \ F_i \equiv \mathcal{N}(\alpha + \beta t_i, \sigma^2), \ i = 1, \ldots, n\}$ and we say $\mathbf{X} \sim P_F \in \mathcal{P}$ if the distribution of the independent Y's is given by $F = (F_1, \ldots, F_n) \in \mathcal{F}$, i.e., $Y_i \sim F_i$.

Example 4: $\mathbf{X} = \{(U_1, V_1), \ldots, (U_n, V_n)\}$ is a random sample of size n from a bivariate normal population with means μ_1 and μ_2, standard deviations $\sigma_1 > 0$ and $\sigma_2 > 0$ and correlation coefficient $\rho \in (-1, 1)$, i.e., we write $(U_i, V_i) \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Let $\mathcal{F}$ be the family of all such bivariate normal distributions and let $\mathcal{P} = \{P_F : \ F \in \mathcal{F}\}$. We say that $\mathbf{X} = \{(U_1, V_1), \ldots, (U_n, V_n)\}$ was generated by P_F (and write $\mathbf{X} \sim P_F$) if (U_1, V_1), ..., (U_n, V_n) are independent, each having the same bivariate normal distribution function F.

The first example is of nonparametric character, because the parameter F that indexes the various $P_F \in \mathcal{P}$ cannot be fit into some finite dimensional space. Also, we deal here with a pure random sample, i.e., with i.i.d. random variables.
The second, third, and fourth example are of parametric nature, since there is a one to one correspondence between F and $\theta = (\mu, \sigma)$ in Example 2, between F and $\theta = (\alpha, \beta, \sigma)$ in Example 3, and between F and $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ in Example 4. We could as well have indexed the possible probability mechanisms by θ, i.e., write P_θ, with θ varying over some appropriate subset $\Theta \subset R^2$, $\Theta \subset R^3$, or $\Theta \subset R^5$, respectively. In Example 3 the data are independent but not identically distributed, since the mean of Y_i changes linearly with t_i.
Of course, we could identify θ with F also in the first example and write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $\Theta = \mathcal{F}$ being of infinite dimensionality in that case. Because of this we will use the same notation describing any family $\mathcal{P}$, namely

$$\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$$

and the whole distinction of nonparametric and parametric probability model disappears in the background, where it is governed by the character of the indexing set Θ.
Many statistical analyses concern themselves with estimating θ, i.e., with estimating the probability mechanism that generated the data. We will assume that we are always able to find such estimates and we denote a generic estimate of θ by $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$, where the emphasized dependence on $\mathbf{X}$ should make clear that any reasonable estimation procedure should be based on the data at hand. Similarly, if we want to emphasize an estimate of P we write $\widehat{P} = P_{\widehat{\theta}}$. Finding any estimate at all can at times be a big order, but that difficulty is not adressed here.
In Example 1 we may estimate $\theta = F$ by the empirical distribution function of the data, i.e, by

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \ I_{(-\infty, x]}(X_i)$$

and we write $\widehat{\theta} = \widehat{F}$. Here $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. Thus $\widehat{F}(x)$ is that fraction of the sample which does not exceed x. $\widehat{F}(x)$ can also be viewed as the cumulative distribution function of a probability distribution which places probability mass $1/n$ at each of the X_i. If some of the X_i coincide, then that common value will receive the appropriate multiple mass.

Often one is content with estimating a particular functional $\psi = \psi(P)$ of P. This will be the situation on which we will focus from now on. A natural estimate of ψ would then be obtained in $\widehat{\psi} = \psi(\widehat{P}) = \psi(P_{\widehat{\theta}})$.
In Example 1 one may be interested in estimating the mean of the sampled distribution F. Then

$$\psi = \psi(P_F) = \int_{-\infty}^{\infty} x \, dF(x)$$

and we obtain

$$\widehat{\psi} = \psi\left(P_{\widehat{F}}\right) = \int_{-\infty}^{\infty} x \, d\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \ ,$$

i.e., the sample average as our estimate of ψ.
Since we index $\mathcal{P}$ by θ we may also speak of estimating a functional $\psi(\theta)$. A natural estimate of $\psi(\theta)$ would then be $\widehat{\psi} = \psi(\widehat{\theta})$. This dual use of $\psi(\theta)$ and $\psi(P)$ should not be confusing, if we keep in mind the convention $\psi(\theta) \equiv \psi(P_\theta)$.
Actually, this functional approach contains the estimation of the full probability model as a special case by using $\psi(\theta) = \theta$. In that case the value set of ψ may be quite large depending on the nature of Θ. However, in most cases we will focus on real valued functionals $\psi(\theta)$.
Having obtained an estimate $\widehat{\psi}$ raises the following questions: How good is it? What is its bias? To what extent does the uncertainty in the original data set influence the estimate, i.e., can we get confidence bounds for the unknown ψ? These are some of the concerns that the bootstrap method tries to address.

1.3 Bootstrap Samples and Bootstrap Distribution

If we had the luxury of knowing θ we could generate B resampled data sets $\mathbf{X}_1, \ldots, \mathbf{X}_B$ from P_θ. For each such data set we could get the corresponding estimate, i.e., obtain $\widehat{\psi}_1, \ldots, \widehat{\psi}_B$. By resampled data set we mean that P_θ generates independent replicates $\mathbf{X}_1, \ldots, \mathbf{X}_B$ just as P_θ generated the original data set $\mathbf{X}$. Of course, it is assumed here that it is always possible to generate such data sets $\mathbf{X}_i$ from P_θ for any given θ. All nonrandom aspects, such as sample sizes within each data set and the values $t_1, \ldots, t_n$ in Example 3, are kept fixed. This should all be understood in the description of the probability model P_θ.

The scatter of these estimates $\widehat{\psi}_1, \ldots, \widehat{\psi}_B$ would be a reflection of the sampling uncertainty in our original estimate $\widehat{\psi}$. As $B \to \infty$, the distribution of the $\widehat{\psi}_1, \ldots, \widehat{\psi}_B$ represents the *sampling distribution* of $\widehat{\psi}$, i.e., we would then be in a position to evaluate probabilities such as

$$Q_A(\theta) = P_\theta(\widehat{\psi} \in A)$$

for all appropriate sets A. This follows from the law of large numbers (LLN), namely

$$\widehat{Q}_A(\theta) = \frac{1}{B} \sum_{i=1}^{B} I_A(\widehat{\psi}_i) \quad \longrightarrow \quad Q_A(\theta)$$

as $B \to \infty$. This convergence is "in probability" or "almost surely" and we will not dwell on it further. Since computing power is cheap, we can let B be quite large and thus get a fairly accurate approximation of $Q_A(\theta)$ by using $\widehat{Q}_A(\theta)$.

Knowledge of this sampling distribution could then be used to set error limits on our estimate $\widehat{\psi}$. For example, we could, by trial and error, find Δ_1 and Δ_2 such that

$$.95 = P_\theta(\Delta_1 \leq \widehat{\psi} \leq \Delta_2) \ ,$$

i.e., 95% of the time we would expect $\widehat{\psi}$ to fall between Δ_1 and Δ_2. This still does not express how far $\widehat{\psi}$ is from the true ψ. This can only be judged if we relate the position of the Δ_i to that of ψ, i.e., write $\delta_1 = \psi - \Delta_1$ and $\delta_2 = \Delta_2 - \psi$ and thus

$$.95 = P_\theta(\widehat{\psi} - \delta_2 \leq \psi \leq \widehat{\psi} + \delta_1) \ .$$

All the above is hypothetical, since in reality we don't know θ. If we did, we would simply evaluate $\psi = \psi(\theta)$ and be done, i.e., we would have no need for estimating it and would have no need for confidence bounds for ψ.

What we know instead, is an estimate $\widehat{\theta}$ of θ. Thus we use $\widehat{P} = P_{\widehat{\theta}}$ when generating B independent replicate data sets $\mathbf{X}_1^\star$, ..., $\mathbf{X}_B^\star$. This collection of alternate data sets is called the *bootstrap sample*. The asterisk on the $\mathbf{X}_j^\star$ emphasizes that these data sets come from $P_{\widehat{\theta}}$ and not from P_θ. Note that $P_{\widehat{\theta}}$ represents a conditional distribution of $\mathbf{X}^\star$ given the original data set $\mathbf{X}$, since $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ is kept fixed in the resampling process. It is as though we treat $\widehat{\theta}$ as the truth, i.e., $P_{\widehat{\theta}}$ as the true probability model which generates the data set $\mathbf{X}^\star$.

For each $\mathbf{X}_i^\star$ obtain the corresponding estimate $\widehat{\theta}_i^\star$ and evaluate $\widehat{\psi}_i^\star = \psi(\widehat{\theta}_i^\star)$. The bootstrap idea is founded in the hope that the scatter of these $\widehat{\psi}_1^\star$, ..., $\widehat{\psi}_B^\star$ should serve as a reasonable proxy for the scatter of $\widehat{\psi}_1, \ldots, \widehat{\psi}_B$ which we cannot observe. If we let $B \to \infty$, we would by the LLN be able to evaluate

$$Q_A(\widehat{\theta}) = \widehat{P}(\widehat{\psi}^\star \in A) = P_{\widehat{\theta}}(\widehat{\psi}^\star \in A)$$

for all appropriate sets A. This evaluation can be done to any desired degree of accuracy by choosing B large enough in our simulations, since

$$\widehat{Q}_A(\widehat{\theta}) = \frac{1}{B}\sum_{i=1}^{B} I_A(\widehat{\psi}_i^\star) \quad \longrightarrow \quad Q_A(\widehat{\theta})$$

as $B \to \infty$. This collection of probabilities is called the *bootstrap distribution* of $\widehat{\psi}^\star$.

In this context and in all future bootstrap appeals to the $WLLN$, it is worth noting that there is a certain similarity between interpreting $\widehat{Q}_A(\widehat{\theta})$ as an approximation of $Q_A(\widehat{\theta})$ for large B and the computation of any analytical result to so many decimal places by some algorithm. Mostly, such computed analytical results are at best approximations. In either case, the more effort one expends, the more accuracy one gets in the approximation. The only real difference between the two is that the simulation approach will not get exactly the same answer when repeated with a different starting seed for the random number generator.

Much of the theoretical bootstrap discussion has focussed on large samples. If the chosen estimate $\widehat{\theta}$ is a reasonable one (namely consistent), then $\widehat{\theta}$ will, in large samples, yield a very good approximation to the unknown θ. Under appropriate continuity conditions, namely

$$P_{\widehat{\theta}} \longrightarrow P_\theta \quad \text{as} \quad \widehat{\theta} \longrightarrow \theta \ ,$$

in a sense, to be left unspecified here, one can then say that the bootstrap distribution of $\widehat{\psi}^\star$ is a good approximation to the sampling distribution of $\widehat{\psi}$, i.e.,

$$P_{\widehat{\theta}}(\widehat{\psi}^\star \in A) \approx P_\theta(\widehat{\psi} \in A) \ .$$

Research has focussed on making this statement more precise by resorting to limit theory. In particular, research has studied the conditions under which this approximation is reasonable and through sophisticated high order asymptotic analysis has tried to reach for conclusions that are meaningful

even for moderately small samples. Our main concern in later sections will be to examine the qualitative behavior of the various bootstrap methods in small samples.

2 The Bootstrap as Bias Reduction Tool

As a first application of the bootstrap method we present its general utility for reducing bias in estimates. The exposition is divided into four subsections. The first covers bias reduction when the functional form of the bias is known. It is pointed out that bias reduction may or may not increase the estimation accuracy, as measured by the mean squared error of the estimate. This is illustrated with two examples. The second subsection shows that the bootstrap can accomplish the same bias reduction without knowing the functional form of the bias. The third subsection discusses the iteration of the bias reduction principle, again assuming a known functional form of the bias. The last subsection shows that this can be accomplished by the iterated bootstrap method without knowing the functional bias form.

2.1 Simple Bias Reduction

Suppose we are interested in estimating a real valued functional $\psi(\theta)$ and we use as estimate $\widehat{\psi} = \psi(\widehat{\theta})$. Such estimates may be biased, i.e., (assuming that expectations are finite)

$$E_\theta\left(\psi(\widehat{\theta})\right) = \psi(\theta) + b(\theta)$$

with bias $b(\theta) \not\equiv 0$. This means that the mean $E_\theta\left(\psi(\widehat{\theta})\right)$ of the $\psi(\widehat{\theta})$ distribution is not centered on the unknown value $\psi(\theta)$, but is off by the bias amount $b(\theta)$.
If we know the functional form of the bias term $b(\theta)$, then the following "bias reduced" estimate

$$\widehat{\psi}_{br1} = \psi(\widehat{\theta}) - b(\widehat{\theta})$$

suggests itself. The subscript 1 indicates that this could be just the first in a sequence of bias reduction iterations, i.e., what we do with $\widehat{\psi}$ for bias reduction we could repeat on $\widehat{\psi}_{br1}$ and so on, see Section 2.3.
Such a correction will typically reduce the bias of the original estimate $\psi(\widehat{\theta})$, but will usually not eliminate it completely, unless of course $b(\widehat{\theta})$ is itself an unbiased estimate of $b(\theta)$.

Note that such bias correction often entails more variability in the bias corrected estimate due to the additional variability of the subtracted bias correction term $b(\widehat{\theta})$. However, it is not clear how the mean squared error of the estimate will be affected by such a bias reduction, since

$$MSE_\theta\left(\widehat{\psi}\right) = E_\theta\left(\widehat{\psi} - \psi\right)^2 = \mathrm{var}_\theta\left(\widehat{\psi}\right) + b^2(\theta) \ .$$

The reduction in bias may well be more than offset by the increase in the variance. In fact, one has the following expression for the difference of the mean squared errors of $\widehat{\psi}_{br1}$ and $\widehat{\psi}$

$$E_\theta\left(\widehat{\psi}_{br1} - \psi\right)^2 - E_\theta\left(\widehat{\psi} - \psi\right)^2 = E_\theta\left(b(\widehat{\theta})^2\right) - 2E_\theta\left(b(\widehat{\theta})(\widehat{\psi} - \psi)\right) \ .$$

There appears to be no obvious way of characterizing the nonnegativity of the right side of this equation, i.e., when bias reduction would lead to increased mean squared error.

As illustration of this point we will present two examples, where the variances increase in both cases and the mean squared errors go in either direction, respectively. In the setting of Example 2 consider first estimating $\psi = \psi(\theta) = \psi(\mu, \sigma) = \sigma^2$. When we use the maximum likelihood estimates

$$\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \ ,$$

we find for $\widehat{\psi} = \psi(\widehat{\theta}) = \widehat{\sigma}^2$

$$E_\theta\left(\widehat{\sigma}^2\right) = \sigma^2 - \frac{\sigma^2}{n} \ ,$$

i.e., the bias is $b(\theta) = -\sigma^2/n$. The bias reduced version is

$$\widehat{\sigma}^2_{br1} = \widehat{\sigma}^2 + \frac{\widehat{\sigma}^2}{n} \ .$$

Here one finds

$$\mathrm{var}_\theta\left(\widehat{\sigma}^2_{br1}\right) = \left(\frac{n+1}{n}\right)^2 \mathrm{var}_\theta\left(\widehat{\sigma}^2\right) > \mathrm{var}_\theta\left(\widehat{\sigma}^2\right)$$

and

$$MSE(\widehat{\sigma}^2) = E_\theta\left(\widehat{\sigma}^2 - \sigma^2\right)^2 = \frac{\sigma^4}{n^4}\left(2n^3 - n^2\right) \ ,$$

$$MSE(\widehat{\sigma}^2_{br1}) = E_\theta\left(\widehat{\sigma}^2_{br1} - \sigma^2\right)^2 = \frac{\sigma^4}{n^4}\left(2(n+1)^2(n-1)+1\right)$$

and thus

$$MSE(\widehat{\sigma}^2) < MSE(\widehat{\sigma}^2_{br1}) \quad \text{for} \quad n > 1\ ,$$

since

$$2(n+1)^2(n-1)+1-(2n^3-n^2) = (3n+1)(n-1) > 0 \ \text{ for } \ n > 1\ .$$

As a second example consider estimating $\psi = \psi(\theta) = \mu^2$ by $\widehat{\psi} = \psi(\widehat{\theta}) = \bar{X}^2$, again in the setting of Example 2. We find

$$E_\theta\left(\bar{X}^2\right) = \mu^2 + \frac{\sigma^2}{n}\ ,$$

i.e., with bias reduced version

$$\widehat{\psi}_{br1} = \bar{X}^2 - \frac{\widehat{\sigma}^2}{n}\ .$$

Here we find

$$\mathrm{var}_\theta\left((\bar{X})^2 - \widehat{\sigma}^2/n\right) = \mathrm{var}_\theta\left((\bar{X})^2\right) + \mathrm{var}_\theta\left(\widehat{\sigma}^2/n\right) > \mathrm{var}_\theta\left((\bar{X})^2\right)$$

and

$$MSE(\widehat{\psi}) = 4\ \frac{\mu^2\sigma^2}{n} + 3\frac{\sigma^4}{n^2}\ ,$$

$$MSE(\widehat{\psi}_{br1}) = 4\ \frac{\mu^2\sigma^2}{n} + \frac{\sigma^4}{n^2}\left(2 + \frac{2n-1}{n^2}\right)$$

and thus clearly

$$MSE(\widehat{\psi}_{br1}) < MSE(\widehat{\psi}) \quad \text{for} \quad n > 1\ ,$$

since

$$3 - \left(2 + \frac{2n-1}{n^2}\right) = \frac{n^2-2n+1}{n^2} = \frac{(n-1)^2}{n^2} > 0 \quad \text{for } \ n > 1\ .$$

2.2 Bootstrap Bias Reduction

In many problems the functional form of the bias term $b(\theta)$ is not known. It turns out that the bootstrap provides us with just the above bias correction without having any knowledge of the function $b(\theta)$. Getting a bootstrap sample of estimates $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$ from $P_{\widehat{\theta}}$ we can form their average

$$\bar{\psi}_B^\star = \frac{1}{B} \sum_{i=1}^{B} \widehat{\psi}_i^\star \ .$$

By the *LLN*

$$\bar{\psi}_B^\star \longrightarrow E_{\widehat{\theta}} \left(\widehat{\psi}^\star \right) = \psi(\widehat{\theta}) + b(\widehat{\theta}) \quad \text{as} \quad B \to \infty \ ,$$

so that $\bar{\psi}_B^\star - \psi(\widehat{\theta})$ is an accurate approximation of $b(\widehat{\theta})$. This can be as accurate as we wish by taking B sufficiently large. Thus we can take as bootstrap bias corrected estimate

$$\widehat{\psi}_{br1}^\star = \psi(\widehat{\theta}) - (\bar{\psi}_B^\star - \psi(\widehat{\theta})) = 2\psi(\widehat{\theta}) - \bar{\psi}_B^\star \ .$$

For large enough B this will be indistinguishable from $\widehat{\psi}_{br1}$, for all practical purposes.

2.3 Iterated Bias Reduction

The bias reduction technique discussed in Section 2.1 can obviously be iterated, as was already hinted in explaining the subscript 1 on $\widehat{\psi}_{br1}$. This works, since $\widehat{\psi}_{br1} = \psi(\widehat{\theta}) - b(\widehat{\theta})$ is again a function of $\widehat{\theta}$ and we thus denote it by $\psi_{br1}(\widehat{\theta})$. Suppose $\widehat{\psi}_{br1}$ is still biased, i.e.,

$$E_\theta \left(\psi_{br1}(\widehat{\theta}) \right) = \psi(\theta) + b_1(\theta) \ .$$

We can also express this as

$$E_\theta \left(\psi_{br1}(\widehat{\theta}) \right) = E_\theta \left(\psi(\widehat{\theta}) - b(\widehat{\theta}) \right) = \psi(\theta) + b(\theta) - E_\theta \left(b(\widehat{\theta}) \right) \ .$$

From these two representations we get

$$b_1(\theta) = - \left\{ E_\theta \left(b(\widehat{\theta}) \right) - b(\theta) \right\} = E_\theta \left(-b(\widehat{\theta}) \right) - (-b(\theta))$$

and thus we can interpret $b_1(\theta)$ as the bias of $-b(\widehat{\theta})$ for estimating $-b(\theta)$. The second order bias reduced estimate thus becomes

$$\begin{aligned}\widehat{\psi}_{br2} = \psi_{br2}(\widehat{\theta}) &= \psi_{br1}(\widehat{\theta}) - b_1(\widehat{\theta}) \\ &= \psi(\widehat{\theta}) - b(\widehat{\theta}) - \left[b(\widehat{\theta}) - E_{\widehat{\theta}}\left(b(\widehat{\theta}^\star)\right)\right] \\ &= \psi(\widehat{\theta}) - 2b(\widehat{\theta}) + E_{\widehat{\theta}}\left(b(\widehat{\theta}^\star)\right) \ ,\end{aligned}$$

where the $\widehat{\theta}^\star$ inside the expectation indicates that its distribution is governed by $\widehat{\theta}$, the subscript on the expectation. Since $\psi_{br2}(\widehat{\theta})$ is a function of $\widehat{\theta}$, we can keep on iterating this scheme and even go to the limit with the iterations. In the two examples of Section 2.1 the respective limits of these iterations result ultimately in unbiased estimates of σ^2 and μ^2, respectively. In the case of the variance estimate the i^{th} iterate gives

$$\begin{aligned}\widehat{\sigma}^2_{bri} &= \widehat{\sigma}^2\left(\frac{1}{n^i} + \frac{1}{n^{i-1}} + \cdots + 1\right) = \widehat{\sigma}^2\ \frac{1 - 1/n^{i+1}}{1 - 1/n} \\ &\to \widehat{\sigma}^2\ \frac{n}{n-1} = s^2 \ \text{ as } \ i \ \to \infty \ ,\end{aligned}$$

where s^2 is the usual unbiased estimate of σ^2. In the case of estimating μ^2 the i^{th} iterate gives

$$\begin{aligned}\widehat{\psi}_{bri} &= \bar{X}^2 - \widehat{\sigma}^2\left(\frac{1}{n} + \cdots + \frac{1}{n^i}\right) = \bar{X}^2 - s^2\ \frac{n-1}{n^2}\ \frac{1 - 1/n^i}{1 - 1/n} \\ &\to \bar{X}^2 - \frac{s^2}{n} \ \text{ as } \ i \to \infty \ ,\end{aligned}$$

the latter being the conventional unbiased estimate of μ^2. In both examples the resulting limiting unbiased estimate is UMVU, i.e., has uniformly minimum variance among all unbiased estimates of the respective target.
According to Hall (1992, p. 32) it is not always clear that these bias reduction iterations should converge to something. He does not give examples. Presumably one may be able to get such examples from situations, in which unbiased estimates do not exist. Since the analysis for such examples is complicated and often involves estimates with infinite expectations, we will not pursue this issue further.

2.4 Iterated Bootstrap Bias Reduction

Here we will examine to what extent one can do the above bias reduction iteration without knowing the forms of the bias functions involved. We will

do this only for the case of one iteration since even that can stretch the simulation capacity of most computers.
Suppose we have generated the i^{th} bootstrap data set $\mathbf{X}_i^\star$ and from it we have obtained $\widehat{\theta}_i^\star$. Then we can spawn a second generation or iterated bootstrap sample $\mathbf{X}_{i1}^{\star\star}, \ldots, \mathbf{X}_{iC}^{\star\star}$ from $P_{\widehat{\theta}_i^\star}$. Each such iterated bootstrap sample then results in corresponding estimates

$$\widehat{\theta}_{i1}^{\star\star},\ \ldots,\ \widehat{\theta}_{iC}^{\star\star}$$

and thus

$$\widehat{\psi}_{i1}^{\star\star},\ \ldots,\ \widehat{\psi}_{iC}^{\star\star}\ , \quad \text{with} \quad \widehat{\psi}_{ij}^{\star\star} = \psi\left(\widehat{\theta}_{ij}^{\star\star}\right)\ .$$

From the LLN we have that

$$\frac{1}{C}\sum_{j=1}^{C}\widehat{\psi}_{ij}^{\star\star} \to E_{\widehat{\theta}_i^\star}\left(\psi(\widehat{\theta}_i^{\star\star})\right) = \psi(\widehat{\theta}_i^\star) + b(\widehat{\theta}_i^\star) \ \text{ as } \ C \to \infty\ .$$

Here $\widehat{\theta}_i^{\star\star}$ inside the expectation varies randomly as governed by $P_{\widehat{\theta}_i^\star}$, while $\widehat{\theta}_i^\star$ is held fixed, just as $\widehat{\theta}^\star$ would vary randomly as governed by $P_{\widehat{\theta}}$, while $\widehat{\theta}$ is held fixed and just as $\widehat{\theta}$ would vary randomly as governed by P_θ, while the true θ is held fixed.
By the LLN and glossing over double limit issues we have that

$$\widehat{A}_{BC} = \frac{1}{B}\sum_{i=1}^{B}\ \frac{1}{C}\sum_{j=1}^{C}\widehat{\psi}_{ij}^{\star\star} \approx \frac{1}{B}\sum_{i=1}^{B}\left(\psi(\widehat{\theta}_i^\star) + b(\widehat{\theta}_i^\star)\right) \to E_{\widehat{\theta}}\left(\psi(\widehat{\theta}^\star) + b(\widehat{\theta}^\star)\right)$$

as $C \to \infty$ and $B \to \infty$. To a good approximation we thus have that

$$\widehat{A}_{BC} \approx E_{\widehat{\theta}}\left(\psi(\widehat{\theta}^\star) + b(\widehat{\theta}^\star)\right) = \psi(\widehat{\theta}) + b(\widehat{\theta}) + E_{\widehat{\theta}}\left(b(\widehat{\theta}^\star)\right)$$

and hence

$$\begin{aligned}
\widehat{\psi}_{br2}^\star &= 3\psi(\widehat{\theta}) - 3\bar{\psi}_B^\star + \widehat{A}_{BC} \\
&\approx 3\psi(\widehat{\theta}) - 3\left(\psi(\widehat{\theta}) + b(\widehat{\theta})\right) + \psi(\widehat{\theta}) + b(\widehat{\theta}) + E_{\widehat{\theta}}\left(b(\widehat{\theta}^\star)\right) \\
&= \psi(\widehat{\theta}) - 2b(\widehat{\theta}) + E_{\widehat{\theta}}\left(b(\widehat{\theta}^\star)\right) = \widehat{\psi}_{br2}\ .
\end{aligned}$$

Note that $\widehat{\psi}_{br2}^\star$ is evaluated completely in terms of $\psi(\widehat{\theta})$, $\widehat{\psi}_i^\star$ and $\widehat{\psi}_{ij}^{\star\star}$, as per definition of $\bar{\psi}_B^\star$ and $\widehat{A}_{BC}$, i.e., without knowledge of the bias functions $b(\cdot)$ and $b_1(\cdot)$.

3 Variance Estimation

Suppose $\mathbf{X} \sim P_\theta$ and we are given an estimate $\widehat{\psi} = \widehat{\psi}(\mathbf{X})$ of the real valued functional $\psi = \psi(\theta)$. We are interested in obtaining an estimate of the variance $\sigma^2_{\widehat{\psi}}(\theta)$ of $\widehat{\psi}$. Such variance estimates are useful in assessing the quality of the estimate $\widehat{\psi}$, especially if the distribution of $\widehat{\psi}$ is approximately normal, as is often the case in large samples. However, such variance estimates are also useful in Studentizing estimates, as for example in the percentile-t bootstrap method of Section 3.4. Here we will briefly mention three general variance estimation procedures. The first is the jackknife method, the second is the substitution method and the third is a bootstrap implementation of the substitution method, that bypasses a major obstacle of the substitution method.

3.1 Jackknife Variance Estimation

When the data vector $\mathbf{X}$ represents a random sample of size n, i.e., $\mathbf{X} = (X_1, \ldots, X_n)$, it often is possible to provide such variance estimates by the jackknife method. See Efron (1982) for a general account. Here we will only briefly indicate the construction of such variance estimates. Let $\widehat{\psi}_{(-i)}$ denote the estimate $\widehat{\psi}$ when it is computed from all observations but the i^{th} one and let

$$\widehat{\psi}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\psi}_{(-i)} \ .$$

Then the jackknife estimate of the variance $\sigma^2_{\widehat{\psi}}(\theta)$ is given by

$$\widehat{\sigma}^2_{\widehat{\psi}J} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\psi}_{(-i)} - \widehat{\psi}_{(\cdot)} \right)^2 \ .$$

Unfortunately, this variance estimate is not always reasonable. For example, if ψ and $\widehat{\psi}$ are population and sample median, respectively, then the above jackknife variance estimate behaves badly in large samples and presumably also in not so large samples, see Efron (1982) for details.
Furthermore, a data vector often has much richer structure than allowed for in a pure random sample scenario. For more complicated structures it is not always clear how to extend the above notion of the jackknife variance estimate.

3.2 Substitution Variance Estimation

Another general variance estimation procedure is based on the following substitution idea. Knowing the functional form of $\sigma^2_{\widehat{\psi}}(\theta)$ (as a function of θ), it would be very natural to simply estimate $\sigma^2_{\widehat{\psi}}(\theta)$ by replacing the unknown parameter θ by $\widehat{\theta}$, namely use as variance estimate

$$\widehat{\sigma}^2_{\widehat{\psi}} = \sigma^2_{\widehat{\psi}}(\widehat{\theta}) \ .$$

Whether $\widehat{\sigma}^2_{\widehat{\psi}}$ itself is a reasonable estimate of $\sigma^2_{\widehat{\psi}}(\theta)$ is another question. In order for this procedure to be reasonable $\sigma^2_{\widehat{\psi}}(\theta)$ needs to be a continuous function of θ, and $\widehat{\theta}$ would have to be a reasonable estimate of θ, i.e., $\widehat{\theta}$ be sufficiently near θ.

3.3 Bootstrap Variance Estimation

The applicability of the above natural substitution procedure is quite general and it can be carried out provided we have the functional form of $\sigma^2_{\widehat{\psi}}(\theta)$ as a function of θ. Unfortunately, this functional form is usually not known. It turns out that the bootstrap method provides a very simple algorithm for getting accurate approximations to $\widehat{\sigma}^2_{\widehat{\psi}}$.

If G_θ denotes the distribution function of $\widehat{\psi}$ with variance $\sigma^2_{\widehat{\psi}}(\theta)$, then $G_{\widehat{\theta}}$ denotes the distribution function of $\widehat{\psi}^\star$ with variance $\sigma^2_{\widehat{\psi}}(\widehat{\theta})$. Here $\widehat{\psi}^\star$ is obtained as estimate from $\mathbf{X}^\star$, which is generated from $P_{\widehat{\theta}}$. In this fashion we can get a bootstrap sample of estimates $\widehat{\psi}^\star_1, \ldots, \widehat{\psi}^\star_B$ and we can compute the sample variance of these bootstrap estimates as

$$\widehat{\sigma}^2_{\widehat{\psi}B} = \frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\psi}^\star_i - \bar{\psi}^\star\right)^2 \ , \quad \text{where} \quad \bar{\psi}^\star = \frac{1}{B}\sum_{i=1}^{B}\widehat{\psi}^\star_i \ .$$

This sample variance is an unbiased estimate of $\sigma^2_{\widehat{\psi}}(\widehat{\theta})$ and its accuracy can be controlled by selecting B sufficiently large, again appealing to the LLN. Thus for all practical purposes we can evaluate the substitution variance estimate $\widehat{\sigma}^2_{\widehat{\psi}}$ by using $\widehat{\sigma}^2_{\widehat{\psi}B}$ instead. Note that this process does not require the functional form of $\sigma^2_{\widehat{\psi}}(\theta)$.

As an illustration we will use Example 1. There consider estimating the mean $\psi(F) = \mu = \int x dF(x)$, using $\widehat{\psi} = \psi(\widehat{F}) = \bar{X}$, with $\widehat{\theta} = \widehat{F}$, the empirical

distribution function of the sample, estimating $\theta = F$. From analytical considerations we know that

$$\sigma^2_{\bar{X}}(F) = \frac{\sigma^2(F)}{n} \ ,$$

where $\sigma(F)$ is the standard deviation of F. The substitution principle would estimate $\sigma^2(F)/n$ by $\sigma^2(\widehat{F})/n$, where

$$\sigma^2\left(\widehat{F}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \ .$$

This $\sigma^2\left(\widehat{F}\right)$ is the variance of $\widehat{F}$, which places probability $1/n$ on each of the X_i, whence the computational formula. Instead of using the analytical form of $\sigma^2_{\bar{X}}(F)$ and substitution, the bootstrap variance estimation method generates B samples, of size n each, from $\widehat{F}$ and computes the B sample averages $\bar{X}^\star_1, \ldots, \bar{X}^\star_B$ of these samples. For large B the sample variance

$$\widehat{\sigma}^2_{\bar{X}B} = \frac{1}{B-1}\sum_{i=1}^{B}\left(\bar{X}^\star_i - \bar{\bar{X}}^\star\right)^2 \ , \quad \text{where} \quad \bar{\bar{X}}^\star = \frac{1}{B}\sum_{i=1}^{B}\bar{X}^\star_i$$

will then be an accurate approximation of $\sigma^2(\widehat{F})/n$. This approximation only requires that we evaluate the averages $\bar{X}^\star_i$ and form their sample variance. No other analytic formula is required in this approach.
By the LLN we again have that $\widehat{\sigma}^2_{\bar{X}B}$ and $\sigma^2(\widehat{F})/n$ are essentially identical for very large B. Of course, here it seems silly to conduct this many simulations and compute the sample variance from such a large bootsstrap sample of estimates, when we could have computed $\sigma^2(\widehat{F})/n$ directly from the original sample. However, this simpler analytic approach is not always available to us, whereas the bootstrap method is applicable universally for variance estimation. The purpose of this example is to show that both approaches reach the same goal.
Here it is worth pointing out that a random sample $\mathbf{X}^\star$ of size n taken from $\widehat{F}$ amounts to sampling n times with replacement from the original sample $X_1, \ldots, X_n$. Since $\widehat{F}$ places probability $1/n$ on each of the X_i, each X_i has the same chance of being selected. Since the resampled observations need to be independent, this sampling from $\{X_1, \ldots, X_n\}$ has to be with replacement.

4 Bootstrap Confidence Bounds

There are many methods for constructing bootstrap confidence bounds. We will not describe them all in detail. The reason for this is that we wish to emphasize the basic simplicity of the bootstrap method and its generality of applicability. Thus we will shy away from any bootstrap modifications which take advantage of analytical devices that are very problem specific and limit the generic applicability of the method.

We will start by introducing Efron's original percentile method, followed by its bias corrected version. The accelerated bias corrected percentile method is not covered as it seems too complicated for general application. It makes use of a certain analytical adjustment, namely the acceleration constant, which is not easily determined from the bootstrap distribution. It is not entirely clear to us whether the method is even well defined in general multiparameter situations not involving maximum likelihood estimates. These three percentile methods are equivariant under monotone transformations on the parameter to be estimated.

Next we will discuss what Hall calls the percentile method and the Student-t percentile method. Finally, we discuss several double bootstrap methods, namely Beran's prepivoting, Loh's calibrated bootstrap, and the automatic double bootstrap. These, but especially the last one, appear to be most promising as far as coverage accuracy in small samples is concerned. However, they also are computationally most intensive. As we go along, we illustrate the methods with specific examples. In a case study we will further illustrate the relative merits of all these methods for small sample sizes in the context of estimating a normal quantile and connect the findings with the approximation rate results given in the literature. All of these investigations concentrate on parametric bootstrap methods, but the definitions are general enough to allow them to be used in the nonparametric context as well. However, in nonparametric settings it typically is not feasible to investigate the small sample coverage properties of the various bootstrap methods, other than by small sample asymptotic methods or by doubly or triply nested simulation loops, the latter being prohibitive. We found that the small sample asymptotics are not very representative of the actual small sample behavior in the parametric case. Thus the small sample asymptotic results in the nonparametric case are of questionable value in really small samples.

Throughout our treatment of confidence intervals, whether by simple bootstrap or by double bootstrap methods, it is often convenient to assume that

the distribution functions F_θ of the estimates $\widehat{\psi}$ are generally continuous and strictly increasing on their support $\{x : 0 < F_\theta(x) < 1\}$. These assumptions allow us to use the probability integral transform result, which states that $U = F_\theta(\widehat{\psi}) \sim U(0,1)$, and quantities like $F_\theta^{-1}(p)$ are well defined without complications. Making this blanket assumption here saves us from repeating it over and over. In some situations it may well be possible to maintain greater validity by arguing more carefully, but that would entail inessential technicalities and distract from getting the basic bootstrap ideas across. It will be up to the reader to perform the necessary detail work, if such generality is desired. If we wish to deviate from the above tacit assumption, we will do so explicitly.

4.1 Efron's Percentile Bootstrap

This method was introduced by Efron (1981). Hall (1992) refers to this also as the "other percentile method," since he reserves the name "percentile method" for another method. In Hall's scheme of viewing the bootstrap Efron's method does not fit in well and he advances various arguments against this "other percentile method." However, he admits that the "other percentile method" performs quite well in the double bootstrap approach. We seem to have found the reason for this as the section on the automatic double bootstrap will make clear. For this reason we prefer not to use the abject term "other percentile method" but instead call it "Efron's percentile method." However, we will usually refer to the percentile method in this section and only make the distinction when confusion with Hall's percentile method is to be avoided. We will first give the method in full generality, present one simple example illustrating what the method does for us, show its transformation equivariance and then provide some justification in the single parameter case.

4.1.1 General Definition

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available the estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. Hence we can obtain a bootstrap sample of estimates $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$ from $P_{\widehat{\theta}}$. The scatter in these bootstrap values should reflect to some degree the uncertainty in our original estimate $\widehat{\psi}$ of ψ. Hence

an appropriately chosen high value of the ordered bootstrap sample

$$\widehat{\psi}^{\star}_{(1)} \leq \ldots \leq \widehat{\psi}^{\star}_{(B)}$$

might serve well as upper confidence bound for ψ. This has some intuitive appeal, but before completely subscribing to this intuition the reader should wait until reading the section on Hall's percentile method.
To make the above definition more precise we appeal to the *LLN*. For sufficiently large B we can treat the empirical distribution of the bootstrap sample of estimates

$$\widehat{G}_B(t) = \frac{1}{B}\sum_{i=1}^{B} I_{[\widehat{\psi}^{\star}_i \leq t]}$$

as a good approximation to the distribution function $G_{\widehat{\theta}}(t)$ of $\widehat{\psi}^{\star}$, where

$$G_{\widehat{\theta}}(t) = P_{\widehat{\theta}}\left(\widehat{\psi}^{\star} \leq t\right) .$$

Solving

$$G_{\widehat{\theta}}(t) = 1-\alpha \quad \text{for} \quad t = \widehat{\psi}_U(1-\alpha) = G^{-1}_{\widehat{\theta}}\left(1-\alpha\right)$$

we will consider $\widehat{\psi}_U(1-\alpha)$ as a nominal $100(1-\alpha)\%$ upper confidence bound for ψ. For large B this upper bound can, for practical purposes, also be obtained by taking $\widehat{G}^{-1}_B(1-\alpha)$ instead of $G^{-1}_{\widehat{\theta}}(1-\alpha)$. This substitution amounts to computing $m = (1-\alpha)B$ and taking the m^{th} of the sorted bootstrap values, $\widehat{\psi}^{\star}_{(1)} \leq \ldots \leq \widehat{\psi}^{\star}_{(B)}$, namely $\widehat{\psi}^{\star}_{(m)}$, as our upper bound. If $m = (1-\alpha)B$ is not an integer, one may have to resort to an interpolation scheme for the two bracketing order statistics $\widehat{\psi}^{\star}_{(k)}$ and $\widehat{\psi}^{\star}_{(k+1)}$, where k is the largest integer $\leq m$. In that case define

$$\widehat{\psi}^{\star}_{(m)} = \widehat{\psi}^{\star}_{(k)} + (m-k)\left(\widehat{\psi}^{\star}_{(k+1)} - \widehat{\psi}^{\star}_{(k)}\right) .$$

When B is sufficiently large, this bootstrap sample order statistic $\widehat{\psi}^{\star}_{(m)}$ is a good approximation of $G^{-1}_{\widehat{\theta}}\left(1-\alpha\right)$. Similarly, one defines

$$\widehat{\psi}_L(\alpha) = G^{-1}_{\widehat{\theta}}\left(\alpha\right) \approx \widehat{G}^{-1}_B(\alpha)$$

as the corresponding nominal $100(1-\alpha)\%$ lower confidence bound for ψ. With $\ell = \alpha B$, it can be obtained as the ℓ^{th} order statistic $\widehat{\psi}^{\star}_{(\ell)}$ of the bootstrap sample of estimates. If ℓ is not an integer, one finds $\widehat{\psi}^{\star}_{(\ell)}$ by interpolation as above.

Together these two bounds comprise a nominal $100(1-2\alpha)\%$, equal tailed confidence interval for ψ. These are the bounds according to Efron's percentile method. The qualifier "nominal" indicates that the actual coverage probabilities of these bounds may be different from the intended or nominal confidence level.
The above construction of upper bound, lower bound, and equal tailed interval shows that generally one only needs to know how to construct an upper bound. At times we will thus only discuss upper or lower bounds.
In situations where we deal with independent, identically distributed data samples, i.e., $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ i.i.d. $\sim F_\theta$, one can show under some regularity conditions that for large sample size n the coverage error is proportional to $1/\sqrt{n}$ for the upper and lower bounds, respectively. Due to fortuitous error cancellation the coverage error is proportional to $1/n$ for the equal tailed confidence interval. What this may really mean in small samples will later be illustrated in some concrete examples.

4.1.2 Example: Bounds for Normal Mean

At this point we will illustrate the method with a very simple example in which the method works very well. The example is presented here to show what the bootstrap method does for us, as compared to analytical methods. Suppose we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a normal population with unknown mean μ, but with known variance σ_0^2. Here the classical $(1-\alpha)\%$ upper confidence bound for the mean μ is obtained as

$$\widehat{\mu}_U(1-\alpha) = \bar{X} + z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} \ ,$$

where $\bar{X}$ is the sample mean and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard normal distribution function Φ. This bound is based on the fact that $\bar{X}$ has a normal distribution with mean μ and variance σ_0^2/n. This is so well known, that it is in the subconscious of most statisticians and one forgets that this is actually an analytical result.
In the bootstrap method we would start with an estimate of the unknown parameter. For simplicity we will take the natural estimate $\widehat{\mu} = \bar{X}$ and will discuss later what would happen if other estimates were chosen. When resampling bootstrap samples $\mathbf{X}_1, \ldots, \mathbf{X}_B$ from $N(\widehat{\mu}, \sigma_0^2)$ and computing the resulting bootstrap sample of estimates

$$(\widehat{\mu}_1^\star, \ldots, \widehat{\mu}_B^\star) = (\bar{X}_1^\star, \ldots, \bar{X}_B^\star) \ ,$$

we know that the empirical distribution function of this sample is a good approximation of

$$G_{\widehat{\mu}}(t) = P_{\widehat{\mu}}\left(\bar{X}^\star \leq t\right) = \Phi\left(\frac{t-\widehat{\mu}}{\sigma_0/\sqrt{n}}\right) ,$$

where the latter equation describes the analytical fact that $\bar{X}^\star \sim N(\widehat{\mu}, \sigma_0^2/n)$, when $\bar{X}^\star$ is the sample mean of $X_1^\star, \ldots, X_n^\star$ i.i.d. $\sim N(\widehat{\mu}, \sigma_0^2)$. The bootstrap method does not know this analytical fact. We only refer to it to see what the bootstrap percentile method generates. The percentile method takes the $(1-\alpha)$-percentile of the bootstrap sample of estimates as upper bound. For large B this percentile is an excellent approximation to $G_{\widehat{\mu}}^{-1}(1-\alpha)$, namely the $(1-\alpha)$-percentile of the $N(\widehat{\mu}, \sigma_0^2/n)$ population or

$$G_{\widehat{\mu}}^{-1}(1-\alpha) = \widehat{\mu} + \Phi^{-1}(1-\alpha)\frac{\sigma_0}{\sqrt{n}} = \widehat{\mu}_U(1-\alpha) .$$

Hence we wind up (approximately) with the classical upper bound just by picking an appropriate percentile of the bootstrap sample of estimates. The analytical results were only used to show that this is the case. They were not used to find the percentile method upper bound. Here the percentile bootstrap method comes up with confidence bounds which have the intended coverage probabilities. This is an accident and is not a general phenomenon, as will be explained in Section 4.1.4. The case where σ^2 is unknown as well is examined later in the context of the bootstrap t-percentile method.

If we had chosen a different estimate for μ, such as the sample median or a trimmed sample mean, there would be no conceptual difference in the application of the percentile bootstrap method. The only thing that would change is that we would compute this type of estimate for each of the resampled samples $\mathbf{X}_i^\star$, $i = 1, \ldots, B$.

Since the sampling distribution of sample mean or trimmed mean is continuous and symmetric around μ we can deduce from the results in Section 4.1.4 that the corresponding percentile bootstrap confidence bounds will have exact coverage rate. When using median or trimmed mean as estimates of μ, the equivalent analytic description of these bounds is complicated and, in the case of the trimmed mean, one has to resort to simulation.

4.1.3 Transformation Equivariance

The property of transformation equivariance is defined as follows. If we have a "method" for constructing confidence bounds for ψ and if $g(\psi) = \tau$ is a strictly increasing transformation of ψ, then we could try to obtain upper confidence bounds for $\tau = \tau(\theta)$ by two methods. On the one hand we can obtain an upper bound $\widehat{\psi}_U$ for ψ and treat $g(\widehat{\psi}_U)$ as upper bound for τ with the same coverage proability, since

$$1 - \alpha = P\left(\widehat{\psi}_U \geq \psi\right) = P\left(g(\widehat{\psi}_U) \geq \tau\right) \ .$$

We refer to this approach as the indirect method. On the other hand we could apply our "method" directly to $\tau = \tau(\theta)$ without reference to ψ, i.e. obtain $\widehat{\tau}_U$. If both applications of our method (direct and indirect) lead to the same result, then we say that the "method" is transformation equivariant. This property is very natural and desirable. It basically says that the method is independent of the way the probability model for the data is parametrized. As it turns out, the percentile method discussed here is transformation equivariant.
The proof of this assertion is based on the identity

$$\tau(\theta) = g(\psi(\theta))$$

and thus on

$$\widehat{\tau}^\star = \tau(\widehat{\theta}^\star) = g(\psi(\widehat{\theta}^\star)) = g(\widehat{\psi}^\star) \ .$$

This in turn implies

$$H_{\widehat{\theta}}(t) = P_{\widehat{\theta}}(\widehat{\tau}^\star \leq t) = P_{\widehat{\theta}}\left(g(\widehat{\psi}^\star) \leq t\right) = P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq g^{-1}(t)\right) = G_{\widehat{\theta}}\left(g^{-1}(t)\right)$$

and thus

$$H^{-1}_{\widehat{\theta}}(p) = g\left(G^{-1}_{\widehat{\theta}}(p)\right) \ .$$

The percentile method applied to $\widehat{\tau}$ yields as upper bound

$$\widehat{\tau}_U = H^{-1}_{\widehat{\theta}}(1-\alpha) = g\left(G^{-1}_{\widehat{\theta}}(1-\alpha)\right) = g\left(\widehat{\psi}_U\right) \ ,$$

i.e., we have the desired transformation equivariance relation between $\widehat{\tau}_U$ and $\widehat{\psi}_U$.

4.1.4 A Justification in the Single Parameter Case

In this subsection we will describe conditions under which the percentile method will give confidence bounds with exact coverage probabilities. In fact, it is shown that the percentile method agrees with the classical bounds in such situations.
Let $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ be an estimate of θ and let $\widehat{\psi} = \psi(\widehat{\theta})$ be the estimate of ψ, the real valued parameter of interest. Consider the situation, in which the distribution of $\widehat{\psi}$ depends only on ψ and not on any other nuisance parameters, although these may be present in the model. Thus we essentially deal with a single parameter problem. Suppose we want to get confidence bounds for $\psi = \psi(\theta)$. Then $\widehat{\psi}$ has distribution function

$$G_\psi(t) = P_\psi\left(\widehat{\psi} \leq t\right) \ .$$

Here we write P_ψ instead of P_θ because of the assumption made concerning the distribution of $\widehat{\psi}$. In order to keep matters simple we will assume that $G_\psi(t)$ is continuous in both t and ψ and that $G_\psi(t) \searrow$ in ψ for fixed t. The latter monotonicity assumption is appropriate for reasonable estimates, i.e., for responsive estimates that tend to increase as the target ψ increases.
Using the probability integral transform we have that $U = G_\psi(\widehat{\psi})$ is distributed uniformly over $[0, 1]$. Thus

$$1 - \alpha = P_\psi\left(G_\psi(\widehat{\psi}) \geq \alpha\right) = P_\psi\left(\psi \leq \widehat{\psi}_{[1-\alpha]}\right)$$

where $\widehat{\psi}_{[1-\alpha]}$ solves

$$G_{\widehat{\psi}_{[1-\alpha]}}(\widehat{\psi}) = \alpha$$

and the above equation results from the equivalence

$$G_\psi(\widehat{\psi}) \geq G_{\widehat{\psi}_{[1-\alpha]}}(\widehat{\psi}) = \alpha \iff \psi \leq \widehat{\psi}_{[1-\alpha]} \ ,$$

invoking the monotonicity of G_ψ in ψ. Hence we can regard $\widehat{\psi}_{[1-\alpha]}$ as a $100(1-\alpha)\%$ upper confidence bound for the parameter ψ.
Now suppose further that there is a monotonically increasing function g and a constant $\tau > 0$ such that

$$\tau\{g(\widehat{\psi}) - g(\psi)\} \sim Z \quad \text{or} \quad g(\widehat{\psi}) \sim g(\psi) + Z/\tau \ ,$$

where Z has a fixed distribution function $H(z)$ which is assumed to be symmetric around 0. This assumption alludes to the fact that sometimes it is

possible to transform estimates in this fashion so that the resulting distribution is approximately standard normal, i.e., Z above would be a standard normal random variable. The consequence of this transformation assumptions is that the percentile method will yield the same upper bound $\widehat{\psi}_{[1-\alpha]}$, and it does so without knowing g, τ or H. Only their existence is assumed in the above transformation.
Under the above assumption we find

$$\begin{aligned} G_\psi(t) = P\left(\widehat{\psi} \leq t\right) &= P\left(\tau\left\{g(\widehat{\psi}) - g(\psi)\right\} \leq \tau\left\{g(t) - g(\psi)\right\}\right) \\ &= H\left(\tau\left\{g(t) - g(\psi)\right\}\right) \ . \end{aligned} \tag{1}$$

Using this identity with $t = \widehat{\psi}$ and $\psi = \widehat{\psi}_{[1-\alpha]}$ we have

$$\alpha = G_{\widehat{\psi}_{[1-\alpha]}}(\widehat{\psi}) = H\left(\tau\left\{g(\widehat{\psi}) - g(\widehat{\psi}_{[1-\alpha]})\right\}\right)$$

and thus

$$\widehat{\psi}_{[1-\alpha]} = g^{-1}\left(g(\widehat{\psi}) - H^{-1}(\alpha)/\tau\right) = g^{-1}\left(g(\widehat{\psi}) + H^{-1}(1-\alpha)/\tau\right) \ ,$$

where the last equality results from the symmetry of H. From Equation (1) we obtain further

$$1 - \alpha = G_\psi\left(G_\psi^{-1}(1-\alpha)\right) = H\left(\tau\left\{g\left(G_\psi^{-1}(1-\alpha)\right) - g(\psi)\right\}\right)$$

and thus

$$G_\psi^{-1}(1-\alpha) = g^{-1}\left(g(\psi) + H^{-1}(1-\alpha)/\tau\right)$$

and replacing ψ by $\widehat{\psi}$ we have

$$G_{\widehat{\psi}}^{-1}(1-\alpha) = g^{-1}\left(g(\widehat{\psi}) + H^{-1}(1-\alpha)/\tau\right) = \widehat{\psi}_{[1-\alpha]} \ .$$

This means that we can obtain the upper confidence bound $\widehat{\psi}_{[1-\alpha]}$ simply by simulating the cumulative distribution function $G_{\widehat{\psi}}(t)$ and then solving $G_{\widehat{\psi}}(t) = 1 - \alpha$ for $t = \widehat{\psi}_{[1-\alpha]}$, i.e., generate a large bootstrap sample of estimates $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$ and for $m = (1-\alpha)B$ take $\widehat{\psi}_{(m)}^\star$, the m^{th} ordered value of $\widehat{\psi}_{(1)}^\star \leq \ldots \leq \widehat{\psi}_{(B)}^\star$, as a good approximation to

$$G_{\widehat{\psi}}^{-1}(1-\alpha) = \widehat{\psi}_{[1-\alpha]} \ .$$

When m is not an integer perform the usual interpolation.

4.2 Bias Corrected Percentile Bootstrap

When our estimate $\widehat{\psi}$ consistently underestimates or overestimates the target ψ it would seem that a bias correction might help matters when setting confidence intervals. This led Efron (1981) to propose also the following bias corrected percentile bootstrap method. It is as easily implemented as the ordinary percentile method and it generally improves matters somewhat. The transformation equivariance property is maintained, but there is a somewhat arbitrary link to the normal distribution. However, for not so small samples a case can often be made that the normal approximation is appropriate when dealing with properly transformed estimates. We give the general definition of the bias corrected percentile method, illustrate its application in the simple example of estimating the normal variance, demonstrate the transformation equivariance, and present an exact coverage justification when the distribution of $\widehat{\psi}$ only depends on ψ and some other normalizing conditions apply.

4.2.1 General Definition

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available an estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. If this estimate satisfies

$$G_\theta(\psi) = P_\theta(\widehat{\psi} \leq \psi) = .5$$

it is called *median unbiased*. For the bootstrap distribution $G_{\widehat{\theta}}$ this entails $G_{\widehat{\theta}}(\widehat{\psi}) = .5$. In order to correct for the bias in estimates that are not median unbiased Efron proposed to compute the following estimated bias correction

$$x_0 = \Phi^{-1}\left(G_{\widehat{\theta}}(\widehat{\psi})\right) \ ,$$

which reduces to zero when $\widehat{\psi}$ is median unbiased. Efron then suggested

$$\widehat{\psi}_{Ubc} = G^{-1}_{\widehat{\theta}}\left(\Phi(2x_0 + z_{1-\alpha})\right)$$

as nominal $(1-\alpha)$-level upper confidence bound for ψ. Here $z_p = \Phi^{-1}(p)$. Similarly,

$$\widehat{\psi}_{Lbc} = G^{-1}_{\widehat{\theta}}\left(\Phi(2x_0 + z_{\alpha})\right)$$

is the corresponding lower bound, i.e., $1-\alpha$ is replaced by α as we go from upper bound to lower bound. Together these two bounds form an equal tailed confidence interval for ψ, with nominal level $(1-2\alpha)\%$. Note that these bounds revert to the Efron percentile bounds when $x_0 = 0$, i.e., when $G_{\widehat{\theta}}(\widehat{\psi}) = .5$.

In practice, one proceeds by obtaining a bootstrap sample of estimates $\widehat{\psi}^\star_1, \ldots, \widehat{\psi}^\star_B$ from $P_{\widehat{\theta}}$ and with $\widehat{q}$ denoting the proportion of these bootstrap values which are $\leq \widehat{\psi}$ one takes $\Phi^{-1}(\widehat{q})$ as a good approximation of x_0. Next determine

$$q_{1-\alpha} = \Phi(2x_0 + z_{1-\alpha}) \ ,$$

compute $m = Bq_{1-\alpha}$ and take the m^{th} ordered value of the $\widehat{\psi}^\star_{(1)} \leq \ldots \leq \widehat{\psi}^\star_{(B)}$, namely $\widehat{\psi}^\star_{(m)}$, as the $(1-\alpha)$-level upper confidence bound for ψ. This then is the upper bound according to the bias corrected percentile method. If m is not an integer, one performs the usual interpolation between the appropriate bracketing order statistics $\widehat{\psi}^\star_{(k)}$ and $\widehat{\psi}^\star_{(k+1)}$. A corresponding procedure is carried out for the lower bound and combining the two bounds results in the usual equal tailed confidence interval. Under certain regularity conditions (see Hall, 1992) one can show, in the i.i.d. case with sample size n, that the coverage error is of order $1/\sqrt{n}$ for either of the bounds and of order $1/n$ for the interval.

4.2.2 Example: Bounds for Normal Variance

In the context of Example 2 we are here interested in confidence bounds for $\psi(\theta) = \psi(\mu, \sigma) = \sigma^2$. As estimates for $\theta = (\mu, \sigma)$ we take the maximum likelihood estimates $\widehat{\mu}$ and $\widehat{\sigma}$. The variance estimate $\widehat{\psi} = \psi(\widehat{\mu}, \widehat{\sigma}) = \widehat{\sigma}^2$ is not median unbiased since

$$G_\theta(\psi) = P_\theta\left(\widehat{\sigma}^2 \leq \sigma^2\right) = P\left(V \leq n\right) = \chi_{n-1}(n) = G_{\widehat{\theta}}(\widehat{\psi}) \ ,$$

where $V = n\widehat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom, with distribution function denoted by $\chi_{n-1}(\cdot)$. The table below illustrates how far from median unbiased $\widehat{\sigma}^2$ is, even for large samples.

n	$\chi_{n-1}(n)$	n	$\chi_{n-1}(n)$	n	$\chi_{n-1}(n)$	n	$\chi_{n-1}(n)$
2	0.843	6	0.694	10	0.650	50	0.567
3	0.777	7	0.679	15	0.622	100	0.547
4	0.739	8	0.667	20	0.605	200	0.533
5	0.713	9	0.658	30	0.586	500	0.521

For the following it is useful to get the distribution function of $\widehat{\psi} = \widehat{\sigma}^2$ explicitly as follows

$$G_\theta(x) = P_\theta(\widehat{\sigma}^2 \leq x) = P(V \leq nx/\sigma^2) = \chi_{n-1}(nx/\sigma^2) \; .$$

Its inverse is

$$G_\theta^{-1}(p) = \chi_{n-1}^{-1}(p) \; \frac{\sigma^2}{n} \; ,$$

where $\chi_{n-1}^{-1}(p)$ is the inverse of χ_{n-1}.
Rather than simulating a bootstrap sample of estimates $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$, we pretend that B is very large, say $B = \infty$, so that we actually have knowledge of the exact bootstrap distribution

$$G_{\widehat{\theta}}(x) = P_{\widehat{\theta}}\left(\widehat{\sigma}^{2\star} \leq x\right) = \chi_{n-1}(nx/\widehat{\sigma}^2) \; .$$

This allows us to write down the bias corrected bootstrap confidence bounds in compact mathematical notation and analyze its coverage properties without resorting to simulations. However, keep in mind that this is not necessary in order to get the bounds. They can always be obtained from the bootstrap sample, as outlined in Section 4.2.1.
The upper confidence bound for $\psi = \sigma^2$ obtained by the bias corrected percentile method can be expressed as

$$\begin{aligned}
\widehat{\psi}_{Ubc} &= G_{\widehat{\theta}}^{-1}\left(\Phi\left(2\Phi^{-1}\left(G_{\widehat{\theta}}(\widehat{\psi})\right) + z_{1-\alpha}\right)\right) \\
&= G_{\widehat{\theta}}^{-1}\left(\Phi\left(2\Phi^{-1}\left(\chi_{n-1}(n)\right) + z_{1-\alpha}\right)\right) \\
&= \chi_{n-1}^{-1}\left(\Phi\left(2\Phi^{-1}\left(\chi_{n-1}(n)\right) + z_{1-\alpha}\right)\right) \; \frac{\widehat{\sigma}^2}{n} \; .
\end{aligned}$$

In comparison, the ordinary Efron percentile upper bound can be expressed as

$$\widehat{\psi}_U = G_{\widehat{\theta}}^{-1}\left(1-\alpha\right) = \chi_{n-1}^{-1}\left(1-\alpha\right) \; \frac{\widehat{\sigma}^2}{n} \; .$$

The actual coverage probabilities of both bounds are given by the following formulas:

$$\begin{aligned}
P_\theta\left(\widehat{\psi}_U \geq \psi\right) &= P_\theta\left(\chi_{n-1}^{-1}(1-\alpha)\widehat{\sigma}^2/n \geq \sigma^2\right) = P(V \geq n^2/\chi_{n-1}^{-1}(1-\alpha)) \\
&= 1 - \chi_{n-1}(n^2/\chi_{n-1}^{-1}(1-\alpha))
\end{aligned}$$

Figure 1: Actual − Nominal Coverage Probability of 95% Upper & Lower Bounds and Asymptotes

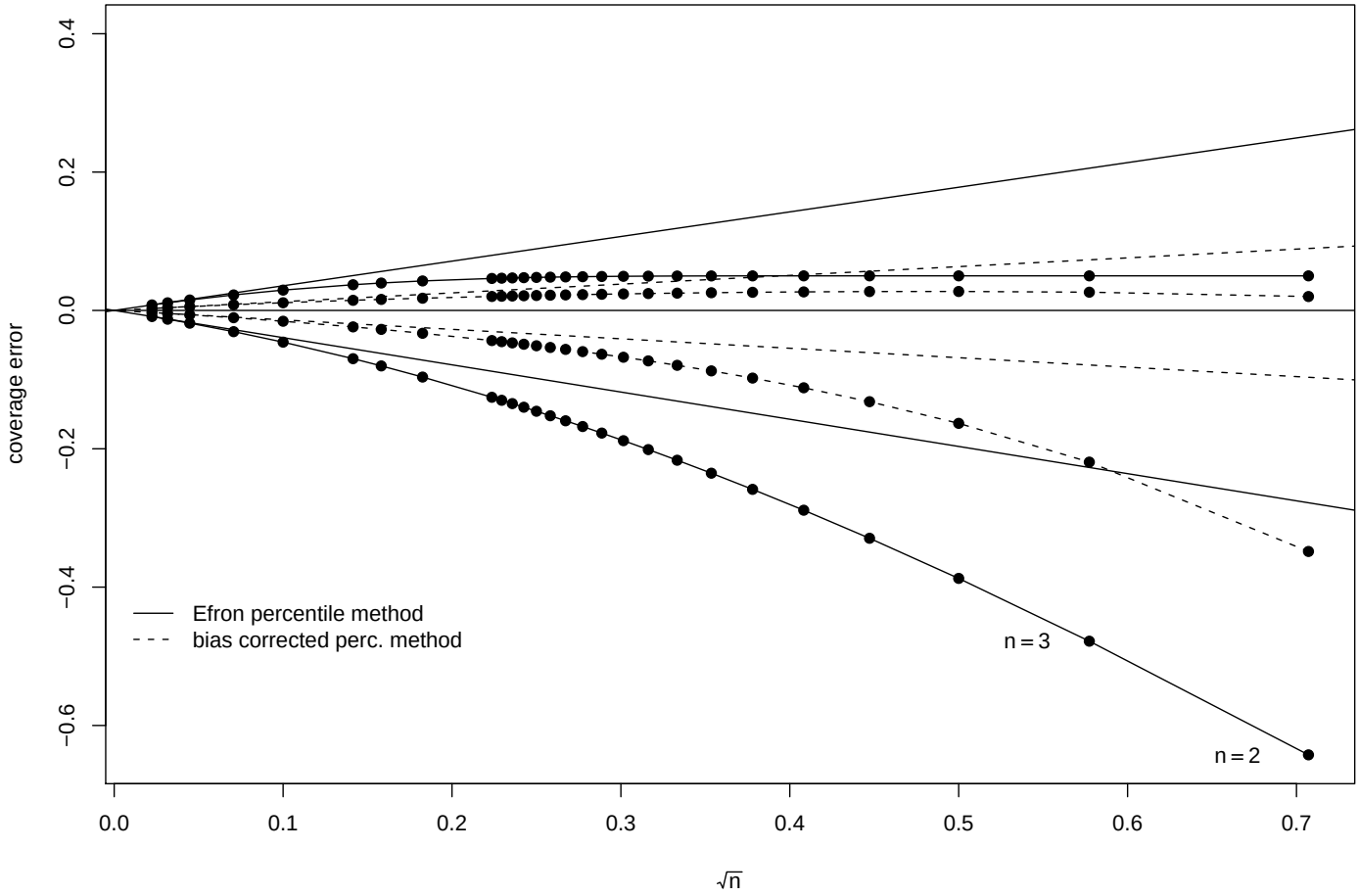
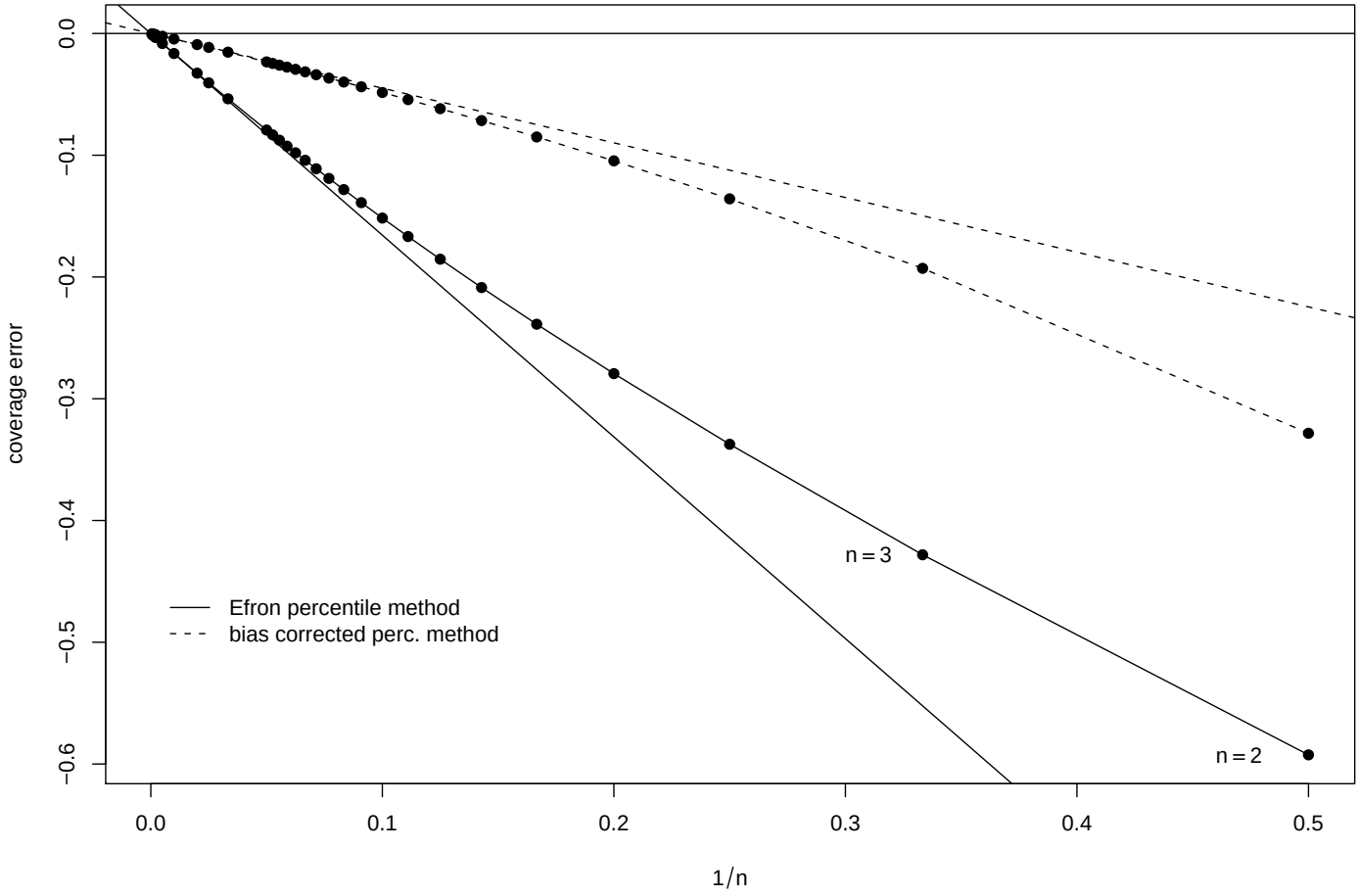


Figure 2: Actual − Nominal Coverage Probability of 90% Confidence Intervals and Asymptotes



and

$$\begin{aligned} P_\theta\left(\widehat{\psi}_{Ubc} \geq \psi\right) &= P_\theta\left(\chi_{n-1}^{-1}\left[\Phi\left(2\Phi^{-1}(\chi_{n-1}(n)) + z_{1-\alpha}\right)\right]\widehat{\sigma}^2/n \geq \sigma^2\right) \\ &= P\left(V \geq n^2/\chi_{n-1}^{-1}\left(\Phi\left[2\Phi^{-1}(\chi_{n-1}(n)) + z_{1-\alpha}\right)\right]\right) \\ &= 1 - \chi_{n-1}\left(n^2/\chi_{n-1}^{-1}\left[\Phi\left(2\Phi^{-1}(\chi_{n-1}(n)) + z_{1-\alpha}\right)\right]\right) \ . \end{aligned}$$

The coverage probabilities for the corresponding lower bounds are the complement of the above probabilities with $1-\alpha$ replaced by α.
Figure 1 shows the coverage error (actual − nominal coverage rate) of nominally 95% upper and lower confidence bounds for σ^2 plotted against the theoretical rate $1/\sqrt{n}$, for sample sizes $n = 2, \ldots, 20, 30, 40, 50, 100, 200, 500,$ 1000, 2000. The asymptotes are estimated by drawing lines through $(0, 0)$ and the points corresponding to $n = 2000$. Note the symmetry of the asymptotes around the zero line, confirming the error cancellation of order $1/\sqrt{n}$. However, the sample size has to be fairly large, say $n \geq 30$, before the asymptotes reasonably approximate the coverage error. The coverage error of the upper bounds is negative and quite substantial for moderate n, whereas that of the lower bounds is positive and small even for moderate n. Throughout, the coverage error of the bias corrected percentile method appears to be smaller than that of the Efron percentile method by a factor of at least two. Figure 2 shows the coverage error (actual − nominal coverage rate) of the corresponding nominally 90% confidence intervals for σ^2 plotted against the theoretical rate of $1/n$. The approximation to the asymptotes is good for much smaller n here. Again the bias corrected version is better by a factor of at least two and for large n by a factor of three.

4.2.3 Transformation Equivariance

Again assume that the parameter of interest is the transform $\tau = g(\psi)$, with g strictly increasing and define $\widehat{\tau} = g(\widehat{\psi})$ as its estimate. The bias corrected percentile method applied directly to the estimate $\widehat{\tau}$ yields as $(1-\alpha)$-level upper bound for τ

$$\widehat{\tau}_{Ubc} = H_{\widehat{\theta}}^{-1}\left(\Phi(2y_0 + z_{1-\alpha})\right)$$

with

$$y_0 = \Phi^{-1}\left(H_{\widehat{\theta}}(\widehat{\tau})\right)$$

and

$$H_{\widehat{\theta}}(t) = P_{\widehat{\theta}}(\widehat{\tau}^\star \leq t) = P_{\widehat{\theta}}\left(g(\widehat{\psi}^\star) \leq t\right) = P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq g^{-1}(t)\right) = G_{\widehat{\theta}}\left(g^{-1}(t)\right) \ .$$

Thus we have

$$y_0 = \Phi^{-1}\left[G_{\widehat{\theta}}\left(g^{-1}\left[g(\widehat{\psi})\right]\right)\right] = \Phi^{-1}\left(G_{\widehat{\theta}}(\widehat{\psi})\right) = x_0$$

and with $H^{-1}_{\widehat{\theta}}(\cdot) = g\left(G^{-1}_{\widehat{\theta}}(\cdot)\right)$ we can write

$$\widehat{\tau}_{Ubc} = g\left(G^{-1}_{\widehat{\theta}}\left(\Phi(2x_0 + z_{1-\alpha})\right)\right) = g\left(\widehat{\psi}_{Ubc}\right) \ ,$$

i.e., the bound has the transformation equivariance property.

4.2.4 A Justification in the Single Parameter Case

Let $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ be an estimate of θ and let $\widehat{\psi} = \psi(\widehat{\theta})$ be the estimate of $\psi = \psi(\theta)$, the real valued parameter for which we desire confidence bounds. Consider again the situation in which the distribution of $\widehat{\psi}$ depends only on ψ and not on any other nuisance parameters, although these may be present in the model. Thus we essentially deal with a single parameter problem. Then $\widehat{\psi}$ has distribution function

$$G_\psi(t) = P_\psi\left(\widehat{\psi} \leq t\right) \ .$$

In order to keep matters simple we will assume that $G_\psi(t)$ is continuous in both t and ψ and that $G_\psi(t) \searrow$ in ψ for fixed t. These are the same assumptions as in Section 4.1.4, where it was shown that this results in exact coverage confidence bounds for ψ. The exact upper confidence bound $\widehat{\psi}_{[1-\alpha]}$ for ψ is found as solution to

$$G_{\widehat{\psi}_{[1-\alpha]}}(\widehat{\psi}) = \alpha \ .$$

Here we assume the existence of an increasing function g and constants $\tau > 0$ and x_0 such that

$$\tau\{g(\widehat{\psi}) - g(\psi)\} + x_0 \sim Z \quad \text{or} \quad g(\widehat{\psi}) \sim g(\psi) - x_0/\tau + Z/\tau \ ,$$

where Z has distribution function $H(z)$, which now is assumed to be standard normal, i.e., $H(z) = \Phi(z)$. Thus, to some extent we have widened the scope over the corresponding assumption in Section 4.1.4 by allowing the bias term x_0, but we also impose the restriction that H has to be standard normal. This restriction may seem severe, but in many situations the distribution

of estimates, transformed in the above fashion, are well approximated by a standard normal distribution. Given the above transformation assumption it is shown below that the bias corrected percentile upper bound for ψ agrees again with $\widehat{\psi}_{[1-\alpha]}$. A priori knowledge of g and τ is not required, they only need to exist. The bias correction constant x_0, which figures explicitly in the definition of the bias corrected percentile method, is already defined in terms of the accessible bootstrap distribution $G_{\widehat{\theta}}(\cdot)$. The remainder of this subsection proves the above claim. The argument is somewhat convoluted and may be skipped.
First we have

$$\begin{aligned} G_\psi(t) &= P_\psi\left(\widehat{\psi} \le t\right) = P_\psi\left(\tau[g(\widehat{\psi}) - g(\psi)] \le \tau[g(t) - g(\psi)]\right) \\ &= P\left(Z \le x_0 + \tau[g(t) - g(\psi)]\right) = \Phi\left(x_0 + \tau[g(t) - g(\psi)]\right) \ . \end{aligned} \quad (2)$$

Replacing (ψ, t) by $(\widehat{\psi}, \widehat{\psi})$ we have

$$G_{\widehat{\psi}}(\widehat{\psi}) = \Phi(x_0) \quad \text{and thus} \quad x_0 = \Phi^{-1}\left(G_{\widehat{\psi}}(\widehat{\psi})\right) \ ,$$

agreeing with the original definition of the bias. The exact upper bound $\widehat{\psi}_{[1-\alpha]}$ is found by solving

$$G_\psi(\widehat{\psi}) = \alpha$$

for ψ. Using Equation (2) for $t = \widehat{\psi}$ and $\psi = \widehat{\psi}_{[1-\alpha]}$ we obtain

$$\alpha = G_\psi(\widehat{\psi}) = \Phi\left(x_0 + \tau[g(\widehat{\psi}) - g(\psi)]\right) \ ,$$

i.e.,

$$z_\alpha = \Phi^{-1}(\alpha) = x_0 + \tau[g(\widehat{\psi}) - g(\psi)]$$

or

$$g(\widehat{\psi}) - g(\psi) = -(x_0 - z_\alpha)/\tau = -(x_0 + z_{1-\alpha})/\tau$$

and finally

$$\widehat{\psi}_{[1-\alpha]} = \psi = g^{-1}\left(g(\widehat{\psi}) + \frac{1}{\tau}(x_0 + z_{1-\alpha})\right) \ . \quad (3)$$

On the other hand, using again Equation (2) (in the second identity below), we have

$$\begin{aligned} \Phi(2x_0 + z_{1-\alpha}) &= G_\psi\left(G_\psi^{-1}\left(\Phi(2x_0 + z_{1-\alpha})\right)\right) \\ &= \Phi\left(x_0 + \tau\left[g\left(G_\psi^{-1}\left[\Phi(2x_0 + z_{1-\alpha})\right]\right) - g(\psi)\right]\right) \ . \end{aligned}$$

Equating the arguments of Φ on both sides we have

$$x_0 + z_{1-\alpha} = \tau \left[g\left(G_{\psi}^{-1} \left[\Phi(2x_0 + z_{1-\alpha}) \right] \right) - g(\psi) \right]$$

or

$$\frac{1}{\tau}(x_0 + z_{1-\alpha}) + g(\psi) = g\left(G_{\psi}^{-1} \left[\Phi(2x_0 + z_{1-\alpha}) \right] \right)$$

and

$$g^{-1}\left(\frac{1}{\tau}(x_0 + z_{1-\alpha}) + g(\psi) \right) = G_{\psi}^{-1} \left[\Phi(2x_0 + z_{1-\alpha}) \right] \ .$$

Replacing ψ by $\widehat{\psi}$ on both sides and recalling Equation (3) we obtain

$$\widehat{\psi}_{[1-\alpha]} = G_{\widehat{\psi}}^{-1} \left[\Phi(2x_0 + z_{1-\alpha}) \right] \ ,$$

i.e., the bias corrected percentile upper bound coincides with the exact upper bound $\widehat{\psi}_{[1-\alpha]}$.

4.3 Hall's Percentile Method

Hall (1992) calls this method simply the percentile method, whereas he refers to Efron's percentile method as "the other percentile method." Using the terms "Efron's percentile method" and "Hall's percentile method" we propose to remove any value judgment and eliminate confusion. It is not clear who first initiated Hall's percentile method, although Efron (1979) already discussed bootstrapping the distribution of $\widehat{\psi} - \psi$, but not in the context of confidence bounds. The method fits well within the general framework that Hall (1992) has built for understanding bootstrap methods. We will first give a direct definition of Hall's percentile method together with its motivation, illustrate it with an example and relate it to Efron's percentile method. The method is generally not transformation equivariant.

4.3.1 General Definition

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available the estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. Instead of bootstrapping the distribution G_θ of $\widehat{\psi}$ we propose here to bootstrap the distribution H_θ of $\widehat{\psi} - \psi$, i.e.,

$$H_\theta(x) = P_\theta(\widehat{\psi} - \psi \leq x) \ .$$

This can be done by simulating a bootstrap sample $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$ and forming

$$\widehat{\psi}_1^\star - \widehat{\psi}, \ldots, \widehat{\psi}_B^\star - \widehat{\psi} \; ,$$

whose empirical distribution function

$$\widehat{H}_B(x) = \frac{1}{B} \sum_{i=1}^{B} I_{[\widehat{\psi}_i^\star - \widehat{\psi} \leq x]} \; ,$$

for large B, approximates

$$H_{\widehat{\theta}}(x) = P_{\widehat{\theta}} \left(\widehat{\psi}^\star - \widehat{\psi} \leq x \right) \; .$$

Here $\widehat{\psi}$ is held fixed within the probability statement $P_{\widehat{\theta}}(\cdots)$ and the term $\widehat{\psi}^\star = \psi(\widehat{\theta}(\mathbf{X}^\star))$ is random with $\mathbf{X}^\star$ generated from the probability model $P_{\widehat{\theta}}$. The bootstrap method here consists of treating $H_{\widehat{\theta}}(x)$ as a good approximation to $H_\theta(x)$, the latter being unknown since it usually depends on the unknown parameter θ. Of course, $\widehat{H}_B(x)$ will serve as our bootstrap approximation to $H_{\widehat{\theta}}(x)$ and thus of $H_\theta(x)$. The accuracy of the first approximation ($\widehat{H}_B(x) \approx H_{\widehat{\theta}}(x)$) can be controlled by the bootstrap sample size B, but the accuracy of $H_{\widehat{\theta}}(x) \approx H_\theta(x)$ depends on the accuracy of $\widehat{\theta}$ as estimate of the unknown θ. The latter accuracy is usually affected by the sample size, which often is governed by other considerations beyond the control of the analyst. Hall's percentile method gives the $100(1-\alpha)\%$ upper confidence bound for ψ as

$$\widehat{\psi}_{HU} = \widehat{\psi} - H_{\widehat{\theta}}^{-1}(\alpha) \; ,$$

and similarly the $100(1-\alpha)\%$ lower confidence bound as

$$\widehat{\psi}_{HL} = \widehat{\psi} - H_{\widehat{\theta}}^{-1}(1-\alpha) \; .$$

The remainder of the discussion will focus on upper bounds, since the discussion for lower bounds would be entirely parallel.
The above upper confidence bound is motivated by the exact $100(1-\alpha)\%$ upper bound

$$\widehat{U} = \widehat{\psi} - H_\theta^{-1}(\alpha) \; ,$$

since

$$\begin{aligned} P_\theta(\widehat{U} > \psi) &= P_\theta \left(\widehat{\psi} - H_\theta^{-1}(\alpha) > \psi \right) \\ &= 1 - P_\theta \left(\widehat{\psi} - \psi \leq H_\theta^{-1}(\alpha) \right) = 1 - H_\theta \left(H_\theta^{-1}(\alpha) \right) = 1 - \alpha \; . \end{aligned}$$

However, $\widehat{U}$ is not a true confidence bound, since it typically depends on the unknown θ through $H_\theta^{-1}(\alpha)$. The bootstrap step consists in sidestepping this problem by approximating $H_\theta^{-1}(\alpha)$ by $H_{\widehat{\theta}}^{-1}(\alpha)$. For large enough B, we can obtain $H_{\widehat{\theta}}^{-1}(\alpha)$ to any accuracy directly from the bootstrap sample of the $D_i = \widehat{\psi}_i^\star - \widehat{\psi}$. Simply order the D_i, i.e., find its order statistics $D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(B)}$ and, for $\ell = B\alpha$, take the ℓ^{th} value $D_{(\ell)}$ as approximation of $H_{\widehat{\theta}}^{-1}(\alpha)$. If ℓ is not an integer interpolate between the appropriate bracketing values of $D_{(k)}$ and $D_{(k+1)}$. Note that it is not required that we know the analytical form of H_θ. All we need to know is how to create new bootstrap samples $\mathbf{X}_i^\star$ from $P_{\widehat{\theta}}$ and thus estimates $\widehat{\psi}_i^\star$ and finally $D_i = \widehat{\psi}_i^\star - \widehat{\psi}$.

In the exceptional case, where $H_\theta^{-1}(\alpha)$ is independent of θ, we have $H_{\widehat{\theta}}^{-1}(\alpha) = H_\theta^{-1}(\alpha) = H^{-1}(\alpha)$ and then the resulting confidence bounds have indeed exact coverage probabilities, if we allow $B \to \infty$.

The basic idea behind this method is to form some kind of pivot, i.e., a function of the data and the parameter of interest, which has a distribution independent of θ. This would be successful if indeed H_θ did not depend on θ. The distribution of $\widehat{\psi}$ will typically depend on θ, but it is hoped that it depends on θ only through $\psi = \psi(\theta)$. Further, it is hoped that this dependence is of a special form, namely that the distribution of $\widehat{\psi}$ depends on ψ only as a location parameter, so that the distribution of $\widehat{\psi} - \psi$ is free of any unknown parameters.

Treating ψ as a location parameter is often justifiable on asymptotic grounds, i.e., for large samples, but may be very misplaced in small samples. In small samples there is really no compelling reason for focussing on the location pivot $\widehat{\psi} - \psi$ as a general paradigm. For example, in the normal variance example discussed earlier and revisited below it would be much more sensible to consider the scale pivot $\widehat{\sigma}^2/\sigma^2$ instead of the location pivot $\widehat{\sigma}^2 - \sigma^2$. Similarly, when dealing with a random sample from the bivariate normal population of Example 4, parametrized by $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ and with the correlation coefficient $\rho = \psi(\theta)$ as the parameter of interest, it would make little sense, except in very large samples, to treat ρ as a location parameter for the maximum likelihood estimate $\widehat{\rho}$.

The focus on $\widehat{\psi} - \psi$ as the proper pivot for Hall's percentile method is mainly justified on asymptotic grounds. The reason for this is that most theoretical bootstrap research has focused on the large sample aspects of the various bootstrap methods.

For other pivots one would have to make appropriate modifications in Hall's percentile method. This is presented quite generally in Beran (1987) and we will illustrate it here with the scale pivot $\widehat{\psi}/\psi$, where it is assumed that the parameter ψ is positive. Suppose the distribution function of $\widehat{\psi}/\psi$ is $H_\theta(x)$ then

$$1-\alpha = P_\theta\left(\widehat{\psi}/\psi > H_\theta^{-1}(\alpha)\right) = P_\theta\left(\psi < \widehat{\psi}/H_\theta^{-1}(\alpha)\right)$$

and replacing the unknown $H_\theta^{-1}(\alpha)$ by $H_{\widehat{\theta}}^{-1}(\alpha)$ gives us the Beran/Hall percentile method upper bound for ψ, namely

$$\widehat{\psi}_{HU} = \widehat{\psi}/H_{\widehat{\theta}}^{-1}(\alpha)\ .$$

From now on, when no further qualifiers are given, it is assumed that a location pivot was chosen in Hall's percentile method. This simplifies matters, especially since it is not always easy to see what kind of pivot would be most appropriate in any given situation, the above normal correlation example being a case in point. Since large sample considerations give some support to location pivots, this default is quite natural.

4.3.2 Example: Bounds for Normal Variances Revisited

Revisiting Example 2, with $\psi = \psi(\theta) = \psi(\mu,\sigma) = \sigma^2$ as parameter of interest, we use again maximum likelihood estimates for $\theta = (\mu,\sigma)$. We are interested in bounds for $\psi(\theta) = \sigma^2$. The distribution function of the location pivot $D = \widehat{\sigma}^2 - \sigma^2$ is

$$\begin{aligned} H_\theta(x) &= P_\theta\left(D \le x\right) = P_\theta\left(\widehat{\sigma}^2 \le x + \sigma^2\right) \\ &= P\left(V \le n + nx/\sigma^2\right) = \chi_{n-1}\left(n + nx/\sigma^2\right)\ . \end{aligned}$$

See Section 3.2.2 for the definition of V and χ_{n-1}. Thus

$$H_\theta^{-1}(\alpha) = \sigma^2\ \left(\frac{\chi_{n-1}^{-1}(\alpha)}{n} - 1\right)$$

and thus

$$H_{\widehat{\theta}}^{-1}(\alpha) = \widehat{\sigma}^2\ \left(\frac{\chi_{n-1}^{-1}(\alpha)}{n} - 1\right)\ ,$$

resulting in the bound

$$\widehat{\sigma}^2_{HU} = \widehat{\sigma}^2 - H_{\widehat{\theta}}^{-1}(\alpha) = \widehat{\sigma}^2\left(2 - \frac{\chi_{n-1}^{-1}(\alpha)}{n}\right)\ .$$

Again we should remind ourselves that this analytic form of $H_{\widehat{\theta}}^{-1}(\alpha)$ is not required in order to compute the upper bound via the Hall percentile method. However, it facilitates the analysis of the coverage rates of the method in this example. This coverage rate can be expressed as

$$\begin{aligned} P_\theta\left(\widehat{\sigma}^2_{HU} \geq \sigma^2\right) &= P_\theta\left(\widehat{\sigma}^2\left(2 - \frac{\chi^{-1}_{n-1}(\alpha)}{n}\right) \geq \sigma^2\right) \\ &= P\left(V \geq \frac{n^2}{2n - \chi^{-1}_{n-1}(\alpha)}\right) = 1 - \chi_{n-1}\left(\frac{n^2}{2n - \chi^{-1}_{n-1}(\alpha)}\right) . \end{aligned}$$

Figure 1a is a repeat of Figure 1 (without asymptotes) with the coverage error of the Hall percentile method added. There is little difference between the Hall and Efron percentile methods in this particular example. Note that the rate is again of order $1/\sqrt{n}$, which happens quite generally under regularity conditions, see Hall (1992). The coverage error rates of the corresponding confidence intervals, not shown here, are again of order $1/n$.

4.3.3 Relation to Efron's Percentile Method

In this subsection we will show that the two percentile methods (Efron's and Hall's) agree when the sampling distribution $G_\theta(x)$ of $\widehat{\psi}$ is continuous and symmetric around ψ, i.e., when

$$G_\theta(\psi + x) = 1 - G_\theta(\psi - x) \quad \text{for all} \;\; x.$$

In terms of the sampling distribution $H_\theta(x)$ of $\widehat{\psi} - \psi$ this symmetry condition is expressed as

$$H_\theta(x) = 1 - H_\theta(-x) \quad \text{for all} \;\; x.$$

The relationship between H_θ and G_θ is the key to the equivalence of the two percentile methods. Namely, we have

$$H_\theta(x) = P_\theta(\widehat{\psi} - \psi \leq x) = P_\theta(\widehat{\psi} \leq x + \psi) = G_\theta(x + \psi) \; .$$

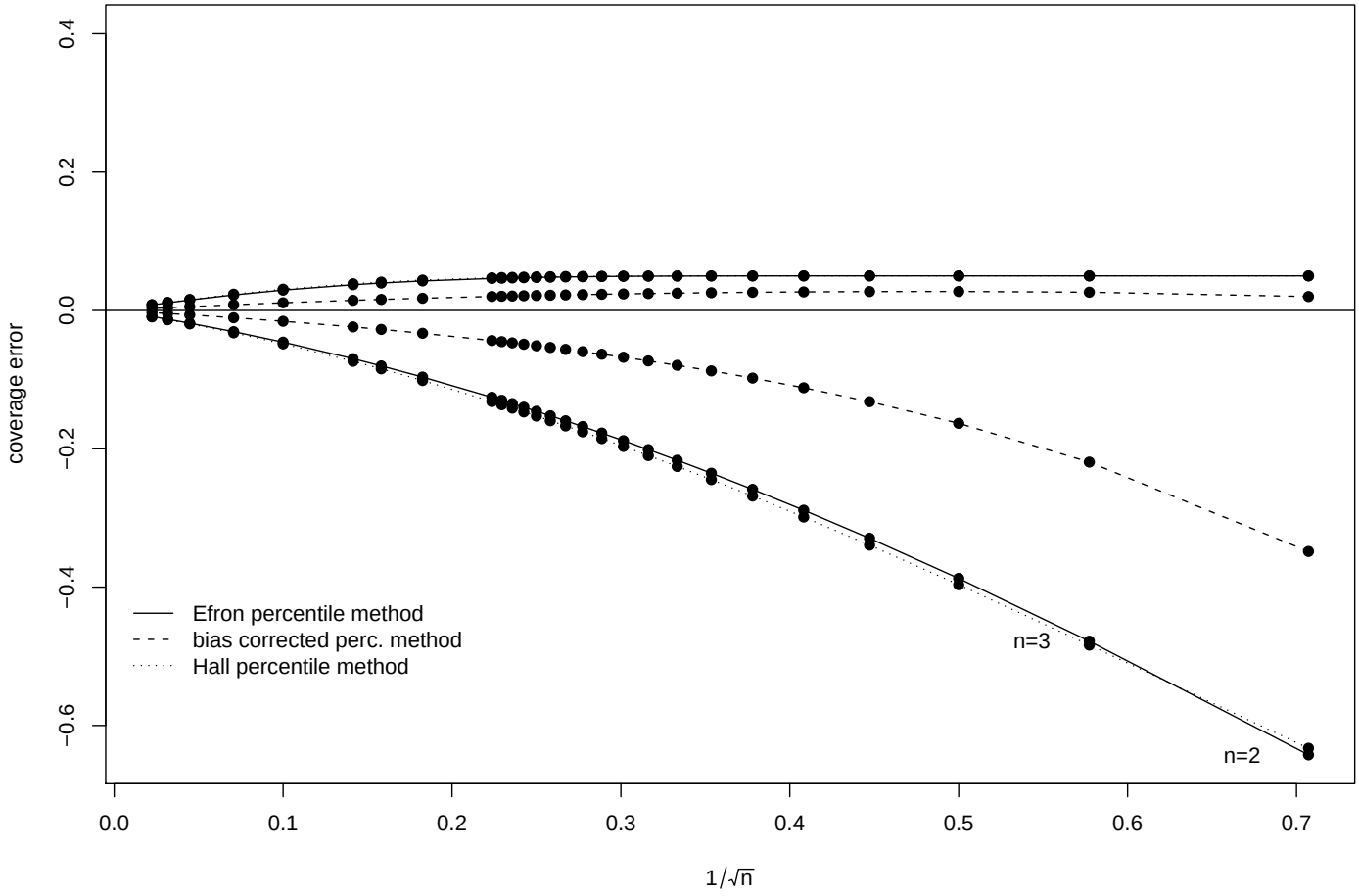
Solving

$$1 - \alpha = H_{\widehat{\theta}}(x) = G_{\widehat{\theta}}(x + \widehat{\psi})$$

for x we get the following two representations for this $(1-\alpha)$-quantile $x = x_{1-\alpha}$:

$$x_{1-\alpha} = H^{-1}_{\widehat{\theta}}(1-\alpha) = G^{-1}_{\widehat{\theta}}(1-\alpha) - \widehat{\psi} \; .$$

Figure 1a: Actual – Nominal Coverage Probability of 95% Upper & Lower Bounds



Hall's percentile upper bound is

$$\widehat{\psi}_{HU} = \widehat{\psi} - H_{\widehat{\theta}}^{-1}(\alpha) = \widehat{\psi} + H_{\widehat{\theta}}^{-1}(1-\alpha) \ ,$$

where the second equality results from the assumed symmetry of $H_{\widehat{\theta}}$. Making use of the dual representation of the above $x_{1-\alpha}$ we find

$$\widehat{\psi}_{HU} = \widehat{\psi} + G_{\widehat{\theta}}^{-1}(1-\alpha) - \widehat{\psi} = G_{\widehat{\theta}}^{-1}(1-\alpha) \ ,$$

which is nothing but Efron's percentile method upper bound.

4.4 Percentile-t Bootstrap

In this subsection we discuss the *percentile-t bootstrap* method for constructing confidence bounds and intervals. It appears that the method was first introduced by Efron (1981). The method is motivated by revisiting the example of confidence bounds for the normal mean, covered in Section 4.1.2 under the assumption of a known variance. This is followed by a general definition and some comments.

4.4.1 Motivating Example

Before giving a definition of the percentile-t method we revisit the example in Section 4.1.2. This time we will assume that both mean and standard deviation of the sampled normal population are unknown and are estimated by the maximum likelihood estimates $\widehat{\mu}$ and $\widehat{\sigma}$. If we were to apply Efron's percentile method to obtain the $(1-\alpha)$-level upper confidence bound for the mean μ, we would be taking the $(1-\alpha)$-quantile of a large bootstrap sample of estimates

$$(\widehat{\mu}_1^\star, \ldots, \widehat{\mu}_B^\star) = (\bar{X}_1^\star, \ldots, \bar{X}_B^\star) \ .$$

These are obtained from bootstrap samples $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ generated from the $N(\widehat{\mu}, \widehat{\sigma}^2)$ distribution. This $(1-\alpha)$-quantile is obtained as the m^{th} value $\bar{X}_{(m)}^\star$ among the ordered bootstrap sample of estimates

$$\bar{X}_{(1)}^\star \leq \ldots \leq \bar{X}_{(B)}^\star \ ,$$

where $m = (1-\alpha)B$. If m is not an integer one performs the usual interpolation. For large B this bound approximately equals the $(1-\alpha)$-quantile

of the bootstrap distribution of $\bar{X}^\star$. This distribution is $N(\widehat{\mu}, \widehat{\sigma}^2/n)$ and its $(1-\alpha)$-quantile is

$$\widehat{\mu}_{zU}(1-\alpha) = \bar{X} + z_{1-\alpha}\frac{\widehat{\sigma}}{\sqrt{n}} \quad \text{with} \quad z_{1-\alpha} = \Phi^{-1}(1-\alpha).$$

Hence Efron's percentile method results in the same bound as in Section 3.1.2 with the only difference being that the previously assumed known σ_0 is replaced by the estimate $\widehat{\sigma}$. The multiplier $z_{1-\alpha}$ remains unchanged. Compare this with the classical upper confidence bound given by

$$\widehat{\mu}_{tU}(1-\alpha) = \bar{X} + t_{n-1}(1-\alpha)\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\frac{n}{n-1}} = \bar{X} + t_{n-1}(1-\alpha)\frac{\widehat{s}}{\sqrt{n}}\,,$$

where $t_{n-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the Student t-distribution with $n-1$ degrees of freedom. This t factor, together with the factor $\sqrt{n/(n-1)}$, adjusts for the sampling variability of the estimate $\widehat{\sigma}$ and results in exact coverage probability for any sample size $n \geq 2$.
In this particular example Hall's percentile method agrees with Efron's, because the sampling distribution of $\widehat{\mu} = \bar{X}$ is continuous and symmetric around μ, see Section 3.3.3. In motivating the transition to the percentile-t method we repeat the derivation in this specific case. Recall that in Hall's percentile method we appeal to the bootstrap distribution $H_{\widehat{\theta}}$ (with $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma})$) of

$$\widehat{\mu}^\star - \widehat{\mu} = \bar{X}^\star - \bar{X}$$

as an approximation to the sampling distribution H_θ of $\widehat{\mu} - \mu = \bar{X} - \mu$. According to Hall's percentile method, the $(1-\alpha)$-level upper bound is obtained by taking the α-quantile of $H_{\widehat{\theta}}$ and forming

$$\widehat{\mu}_{HU}(1-\alpha) = \widehat{\mu} - H^{-1}_{\widehat{\theta}}(\alpha) = \bar{X} - z_\alpha\frac{\widehat{\sigma}}{\sqrt{n}} = \bar{X} + z_{1-\alpha}\frac{\widehat{\sigma}}{\sqrt{n}} = \widehat{\mu}_{zU}(1-\alpha)\,.$$

Bootstrapping the distribution of $\bar{X}^\star$ or $\bar{X}^\star - \bar{X}$ certainly mimics the sampling variability of $\bar{X}$ relative to μ, but it does not capture the sampling variability of the estimate $\widehat{\sigma}$, which explicitly is part of the formula for $\widehat{\mu}_{HU} = \widehat{\mu}_{zU}$. Note that the percentile method (Hall's or Efron's) uses $\widehat{\sigma}$ only to obtain samples from $N(\widehat{\mu}, \widehat{\sigma}^2)$ and does not use the above formula to obtain $\widehat{\mu}_{HU} = \widehat{\mu}_{zU}$. However, the formula is useful in showing explicitly what either of the two percentile methods accomplishes in this example. Namely, the z-factor in

the above formula indicates, that the percentile methods act as though $\widehat{\sigma}$ is equal to the true (unknown) standard deviation σ, in which case the use of the z-factor would be most appropriate. Since $\widehat{\sigma}$ varies around σ from sample to sample, this sampling variation needs to be accounted for in setting confidence bounds.

The percentile-t method carries the pivoting step of Hall's percentile method (of bootstrapping $\bar{X} - \mu$) one step further by considering a Studentized pivot

$$T = \frac{\bar{X} - \mu}{\widehat{\sigma}} \ .$$

If we knew the distribution function $K_\theta(x)$ of T we could obtain a $(1-\alpha)$ level upper confidence bound for μ as follows:

$$\bar{X} - K_\theta^{-1}(\alpha)\widehat{\sigma}$$

since

$$1 - \alpha = P\left(\frac{\bar{X} - \mu}{\widehat{\sigma}} \geq K_\theta^{-1}(\alpha)\right) = P\left(\bar{X} - K_\theta^{-1}(\alpha)\widehat{\sigma} \geq \mu\right) \ .$$

The subscript θ on $K_\theta^{-1}(\alpha)$ allows for the possibility that the distribution of T may still depend on θ. In this particular example K is independent of θ and thus $\bar{X} - K^{-1}(\alpha)\widehat{\sigma}$ is an exact $(1-\alpha)$-level upper confidence bound for μ. To obtain $K^{-1}(\alpha)$ we can either appeal to tables of the Student-t distribution, because for this particular example we know that

$$K^{-1}(\alpha) = t_{n-1}(\alpha)\widehat{\sigma}/\sqrt{n-1} = -t_{n-1}(1-\alpha)\widehat{\sigma}/\sqrt{n-1} \ ,$$

or, in a more generic approach, we can simulate the distribution K of T by generating samples from $N(\mu, \sigma^2)$ for any $\theta = (\mu, \sigma)$, since in this example K is not sensitive to the choice of θ. However, for reasons to be explained in the next section, we may as well simulate independent samples $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ from $N(\widehat{\mu}, \widehat{\sigma}^2)$ and generate $T_1^\star, \ldots, T_B^\star$ with $T_i^\star = (\bar{X}_i^\star - \bar{X})/\widehat{\sigma}_i^\star$ computed from the i^{th} bootstrap sample $\mathbf{X}_i^\star$. For very large B this simulation process will approximate the bootstrap distribution $\widehat{K} = K$ of

$$T^\star = \frac{\bar{X}^\star - \bar{X}}{\widehat{\sigma}^\star} \ .$$

The percentile-t method constructs the $(1-\alpha)$-level upper confidence bound as

$$\widehat{\mu}_{tU} = \bar{X} - \widehat{K}^{-1}(\alpha)\widehat{\sigma} \ .$$

For $\ell = \alpha B$ we can consider the ℓ^{th} ordered value of $T^\star_{(1)} \leq \ldots \leq T^\star_{(B)}$, namely $T^\star_{(\ell)}$, as an excellent approximation to $\widehat{K}^{-1}(\alpha)$. When αB is not an integer one does the usual interpolation of the appropriate adjacent ordered values $T^\star_{(k)}$ and $T^\star_{(k+1)}$.
By bootstrapping the distribution of the Studentized ratio T we hope that we capture to a large extent the sampling variability of the scale estimate used in the denominator of T. That this may not be completely successful is reflected in the possiblity that the distribution K_θ of T may still depend on θ.
The above discussion gives rise to a small excursion, which is not an integral part of the percentile-t method, but represents a rough substitute for it. Since $\bar{X}^\star_{(m)} \approx \widehat{\mu}_{zU}(1-\alpha)$, Efron (1982) considered the following t-factor patch to the Efron percentile method, namely

$$\bar{X} + \frac{t_{n-1}(1-\alpha)}{z_{1-\alpha}} \sqrt{\frac{n}{n-1}} \left(\bar{X}^\star_{(m)} - \bar{X} \right) \ ,$$

with $m = (1-\alpha)B$. This patched version of the Efron percentile method upper bound is approximately equal to the above $\widehat{\mu}_{tU}(1-\alpha)$, as is seen from

$$\bar{X}^\star_{(m)} \approx \widehat{\mu}_{zU}(1-\alpha) = \bar{X} + z_{1-\alpha} \frac{\widehat{\sigma}}{\sqrt{n}}$$

$$\Rightarrow \quad \bar{X}^\star_{(m)} - \bar{X} \approx z_{1-\alpha} \frac{\widehat{\sigma}}{\sqrt{n}}$$

$$\Rightarrow \quad \frac{t_{n-1}(1-\alpha)}{z_{1-\alpha}} \sqrt{\frac{n}{n-1}} \left(\bar{X}^\star_{(m)} - \bar{X} \right) \approx t_{n-1}(1-\alpha) \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{\frac{n}{n-1}}$$

and thus

$$\bar{X} + \frac{t_{n-1}(1-\alpha)}{z_{1-\alpha}} \sqrt{\frac{n}{n-1}} \left(\bar{X}^\star_{(m)} - \bar{X} \right) \approx \widehat{\mu}_{tU}(1-\alpha) \ .$$

This idea of patching the Efron percentile method can be applied to other situations as well, especially when estimates are approximately normal. The effect is to widen the bounds in order to roughly protect the coverage confidence. In this particular example the patch works perfectly in that it results in the classical bound. The patch is easily applied, provided we have a reasonable idea of the degrees of freedom to use in the t-factor correction. However, Efron (1982) warns against its indiscriminate use. Note also that in applying the patch we lose the transformation equivariance of Efron's percentile method.

4.4.2 General Definition

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available the estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. Furthermore, it is assumed that we have some scale estimate $\widehat{\sigma}_{\widehat{\psi}}$, so that we can define the Studentized pivot

$$T = \frac{\widehat{\psi} - \psi}{\widehat{\sigma}_{\widehat{\psi}}}.$$

In order for T to be a pivot in the strict sense, its distribution would have to be independent of any unknown parameters. This is not assumed here, but if this distribution K_θ depends on θ, it is hoped that it does so only weakly. The $(1-\alpha)$-level percentile-t upper bound for ψ is defined as

$$\widehat{\psi}_{tU} = \widehat{\psi} - K^{-1}_{\widehat{\theta}}(\alpha)\widehat{\sigma}_{\widehat{\psi}}\,.$$

Here $K^{-1}_{\widehat{\theta}}(\alpha)$ is obtained by simulating the distribution $K_{\widehat{\theta}}$ of

$$T^\star = \frac{\widehat{\psi}^\star - \widehat{\psi}}{\widehat{\sigma}^\star_{\widehat{\psi}}}\,.$$

This is done by simulating samples $\mathbf{X}^\star_1, \ldots, \mathbf{X}^\star_B$ from $P_{\widehat{\theta}}$, generating $T^\star_1, \ldots, T^\star_B$, with

$$T^\star_i = \frac{\widehat{\psi}^\star_i - \widehat{\psi}}{\widehat{\sigma}^\star_{\widehat{\psi}i}}$$

computed from the i^{th} bootstrap sample $\mathbf{X}^\star_i$. For $\ell = \alpha B$ take the ℓ^{th} ordered value $T^\star_{(\ell)}$ of the order statistics

$$T^\star_{(1)} \leq \ldots \leq T^\star_{(B)}$$

as a good approximation of $K^{-1}_{\widehat{\theta}}(\alpha)$. When $\ell = \alpha B$ is not an integer, perform the usual interpolation between the appropriate adjacent order statistics $T^\star_{(k)}$ and $T^\star_{(k+1)}$.

In the definition of the percentile-t upper bound the estimated quantile $K^{-1}_{\widehat{\theta}}(\alpha)$ was used instead of the more appropriate but unknown $K^{-1}_{\theta}(\alpha)$. Replacing the unknown parameter θ by the estimate $\widehat{\theta}$ has two motivations. First, it is practical, because we know $\widehat{\theta}$, and second, $\widehat{\theta}$ is in the vicinity

of the true, but unknown value of θ, and thus $K^{-1}_{\widehat{\theta}}(\alpha)$ is likely to be more relevant than taking any value of θ in $K^{-1}_{\theta}(\alpha)$ and solely appealing to the insensitivity of K_θ with respect to θ.
The above definition of percentile-t bounds is for upper bounds, but by switching from α to $1-\alpha$ we are covering $1-\alpha$ lower bounds as well. Combining $(1-\alpha)$-level upper and lower bounds we obtain $(1-2\alpha)$-level confidence intervals for ψ.

4.4.3 General Comments

The percentile-t bootstrap method appears to have improved coverage properties. In fact, under regularity conditions it can be shown (see Hall, 1988) that the coverage error of one-sided percentile-t bounds is of order $1/n$, in contrast to the $1/\sqrt{n}$ rate in the case of Hall's and Efron's percentile or bias corrected percentile method.
The method shares with Hall's percentile method the drawback that it is generally not transformation equivariant.
Also, Studentization makes most sense when ψ is a location parameter, but that is not always the case. In Example 4, with $\psi = \rho$, we can hardly treat ρ as a location parameter and Studentization has performed poorly here. The z-transform has been suggested as the appropriate cure for this problem, namely applying the percentile-t method to the transformed parameters $z = \frac{1}{2}\log\{(1+\rho)/(1-\rho)\}$ and corresponding estimates. The confidence bounds for z are then backtransformed to confidence bounds for ρ. However, this is a very problem specific fix and not useful as a general bootstrap tool.
A further disadvantage of the percentile-t method is the source of its better coverage properties, namely the explicit requirement of an appropriate scale estimate for Studentization. Such a scale estimate, to serve its purpose, should be distributionally proportional to the standard deviation of the original estimate $\widehat{\psi}$. Section 3 discusses several schemes for getting variance estimates of estimates, of which the bootstrap method is the most versatile. If we do employ bootstrap variance estimates in order to accomplish the Studentization, then that would require an extra level of simulations for each of the $T^\star_i$ to be generated. In effect, this would amount to some form of double bootstrap, which is the topic of discussion in the next section, although not from the percentile-t perspective.

5 Double Bootstrap Confidence Bounds

This section introduces two closely related double bootstrap methods for constructing confidence bounds. Single bootstrapping amounts to generating B bootstrap samples, where B is quite large, typically $B = 1,000$, and computing estimates for each such bootstrap sample. In double bootstrapping each of these B bootstrap samples spawns itself a set of A second order bootstrap samples. Thus, all in all, $A \cdot B$ samples will be generated with the attending data analyses to compute estimates, typically $A \cdot B + B$ of them. If $A = B = 1000$ this amounts to $1,001,000$ such analyses and is thus computationally very intensive. This is a high computational price to pay, especially when the computation of estimates $\widehat{\theta}(\mathbf{X})$ is costly to begin with. If that cost grows with the sample size of $\mathbf{X}$, one may want to limit its use only to analyses involving small sample sizes, but that is the area where coverage improvement makes most sense anyway. Before these methods will be used routinely, progress will need to be made in computational efficiency. We hope that some time soon clever algorithms will be found that reduce the effort of $A \cdot B$ simulations to $k \cdot B$, where k is of the order of ten. Such a reduction would make these double bootstrap methods definitely the preferred choice as a general tool for constructing confidence bounds.

It appears that methods based on double bootstrap approaches are most successful in maintaining the intended coverage rates for the resulting confidence bounds. A first application of the double bootstrap method to confidence bounds surfaced in the last section when discussing the possibility of bootstrap scale estimates to be used in the bootstrapped Studentized pivots of the percentile-t method. Here we first discuss Beran's (1987) method, which is based on the concept of a root (a generalization of the pivot concept) and the prepivoting idea. The latter invokes an estimated probability integral transform in order to obtain improved pivots, which then are bootstrapped. It is shown that Beran's method is equivalent to Loh's (1987) calibration of confidence coefficients. This calibration uses the bootstrap method to estimate the coverage error with the aim of correcting for it. The second iterated bootstrap method, proposed by Scholz (1992), automatically finds the proper natural pivot when such pivots exist. This yields confidence bounds with essentially exact coverage whenever these are possible.

5.1 Prepivot Bootstrap Methods

This subsection introduces the concept of a root, motivates the use of roots by showing how confidence bounds are derived from special types of roots, namely from exact pivots. Then single bootstrap confidence bounds, based on roots, are introduced and seen to be a simple extension of Hall's percentile method. These confidence sets are based on an estimated probability integral transform. This transform can be iterated, which suggests the prepivoting step. The effect of this procedure is examined analytically in a special example, where it results in exact coverage. Since analysis is not always, feasible it is then shown how to accomplish the same by an iterated bootstrap simulation procedure. This is concluded with remarks about the improved large sample properties of the prepivot methods and with some critical comments.

5.1.1 The Root Concept

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available the estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. Beran (1987) introduces the concept of a *root*. This is a function $R = R(\mathbf{X}, \psi) = R(\mathbf{X}, \psi(\theta))$ of θ (through $\psi(\theta)$) and the data $\mathbf{X}$. If the distribution function

$$F_\theta(r) = P_\theta(R \leq r) = P_\theta\left(R(\mathbf{X}, \psi(\theta)) \leq r\right)$$

of such a root does not depend on θ, then R is a pivot in the strict sense. The idea behind pivots is to play off the double dependence on θ in the above probability statement, namely through P_θ and $\psi(\theta)$, so that no dependence on θ remains. Such pivots are not always possible. In fact, they are the exception and not the rule. It was the possible dependence of F_θ on θ, which led Beran to introduce this broader terminology of *root*, and we follow his example at least in this section. Before describing the use of such roots for constructing confidence bounds, we will discuss the procedure in the case of strict pivots.

5.1.2 Confidence Sets From Exact Pivots

Pivots have long been instrumental in finding confidence bounds. In this subsection we will assume that $F_\theta(r) = F(r)$ is independent of θ. Let $r_{1-\alpha}$ be such that $F(r_{1-\alpha}) = 1 - \alpha$ or $r_{1-\alpha} = F^{-1}(1-\alpha)$. Now let

$$C(\mathbf{X}, 1-\alpha) = \{\psi : R(\mathbf{X}, \psi) \leq r_{1-\alpha}\} \ ,$$

then $C(\mathbf{X}, 1-\alpha)$ can be considered a $(1-\alpha)$-level confidence set for ψ. This results from

$$P_\theta\left(\psi \in C(\mathbf{X}, 1-\alpha)\right) = P_\theta\left(R(\mathbf{X}, \psi) \leq r_{1-\alpha}\right) = F(r_{1-\alpha}) = 1-\alpha\ .$$

Typically, when $R(\mathbf{X}, \psi)$ is monotone in ψ, the set $C(\mathbf{X}, 1-\alpha)$ will be some kind of interval, infinite on the right or left, which is equivalent to either a lower or upper confidence bound for ψ. When $R(\mathbf{X}, \psi)$ is first decreasing and then increasing in ψ, we will usually obtain a bounded confidence interval for ψ.

Often the distribution F is known analytically for certain pivots and the quantiles r_α are tabulated. As an example, consider Example 2, where we are interested in confidence bounds for $\psi = \psi(\mu, \sigma) = \mu$. Then

$$R = \sqrt{n}\,\frac{\bar{X}-\psi}{S}\ , \quad \text{with} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\ ,$$

is a pivot with distribution function given by the tabulated Student-t distribution with $n-1$ degrees of freedom. Following the above generic recipe for $C(\mathbf{X}, 1-\alpha)$ with $r_{1-\alpha} = t_{n-1}(1-\alpha)$ we get the following lower confidence bound for ψ

$$\bar{X} - t_{n-1}(1-\alpha)\frac{S}{\sqrt{n}}\ .$$

Upper bounds can be obtained by changing R to $-R$ and intervals can be obtained by changing R to $|R|$.

In some situations the pivot distribution F can not be determined analytically. Then the only recourse is simulation. As an example consider the case where $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample from the extreme value distribution

$$H(x) = 1 - \exp\left(-\exp\left(\frac{x-a}{b}\right)\right) \quad \text{for} \quad -\infty < x < \infty\ ,$$

where $a \in R$ and $b > 0$ are the unknown parameters, i.e., $\theta = (a, b)$. Such a random sample can also be considered as a log-transform of a random sample from a Weibull distribution with scale parameter $\kappa = \exp(a)$ and shape parameter $\beta = 1/b$. If $\widehat{a}$ and $\widehat{b}$ are the maximum likelihood estimates of a and b, one can treat

$$R_1 = \frac{\widehat{a}-a}{\widehat{b}} \quad \text{and} \quad R_2 = \frac{b}{\widehat{b}}$$

as appropriate pivots for a and b, respectively. However, their distribution is not practically obtainable by analytical methods. By simulating R_1 and R_2 for many random samples generated from one specifice extreme value distribution (it does not matter which, because of the pivot property) one can obtain accurate estimates of these pivot distributions. Bain (1978) has tabulated the simulated quantiles of these pivots. In a sense these can be considered as a forerunner of the bootstrap method.

5.1.3 Confidence Sets From Bootstrapped Roots

Here we assume that the distribution F_θ of R may still depend on θ. One can then use the bootstrap approach and estimate F_θ by $F_{\widehat{\theta}}$. We do not need to know the functional form of F_θ, i.e., we don't have to plug an estimate $\widehat{\theta}$ for θ into F_θ in order to obtain $F_{\widehat{\theta}}$.We can instead simulate a bootstrap sample $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ from $P_{\widehat{\theta}}$, in which case we only need to know how to plug in $\widehat{\theta}$ into P_θ to get $P_{\widehat{\theta}}$ and how to generate samples from it. From this bootstrap sample compute the corresponding bootstrap sample of roots

$$\Big(R(\mathbf{X}_1^\star, \widehat{\psi}), \ldots, R(\mathbf{X}_B^\star, \widehat{\psi})\Big) \ ,$$

where $\widehat{\psi} = \psi(\widehat{\theta})$. Note that we have replaced all appearances of θ by $\widehat{\theta}$, i.e., in the distribution $P_{\widehat{\theta}}$ generating the bootstrap samples $\mathbf{X}_i^\star$ and in $\widehat{\psi} = \psi(\widehat{\theta})$. For large B this bootstrap sample of roots will give an accurate description of $F_{\widehat{\theta}}(\cdot)$, namely

$$\frac{1}{B}\sum_{i=1}^{B} I_{[R(\mathbf{X}_i^\star, \widehat{\psi}) \le x]} \longrightarrow F_{\widehat{\theta}}(x) \quad \text{as} \quad B \to \infty \ .$$

By sorting the bootstrap sample of roots we can, by the usual process, get a good approximation to the quantile $r_{1-\alpha}(\widehat{\theta})$, which is defined by

$$F_{\widehat{\theta}}\Big(r_{1-\alpha}(\widehat{\theta})\Big) = 1 - \alpha \quad \text{or} \quad r_{1-\alpha}(\widehat{\theta}) = F_{\widehat{\theta}}^{-1}(1-\alpha) \ .$$

To get a bootstrap confidence set for ψ one replaces $r_{1-\alpha}$ in $C(\mathbf{X}, 1-\alpha)$ by $r_{1-\alpha}(\widehat{\theta})$ or by its just suggested approximation. We will not distinguish between the two. Thus we have the following bootstrap confidence set

$$C_B(\mathbf{X}, 1-\alpha) = \Big\{\psi : R(\mathbf{X}, \psi) \le r_{1-\alpha}(\widehat{\theta})\Big\} = \Big\{\psi : F_{\widehat{\theta}}\big(R(\mathbf{X}, \psi)\big) \le 1-\alpha\Big\} \ .$$

The second representation of C_B shows that the construction of the confidence set appeals to the probability integral transform. Namely, for continuous F_θ the random variable $U = F_\theta(R)$ has a uniform distribution on the interval $(0,1)$ and then $P(U \leq 1-\alpha) = 1-\alpha$. Unfortunately, we can only use the estimated probability integral transform $\widehat{U} = F_{\widehat{\theta}}(R)$ and $\widehat{U}$ is no longer distributed uniformly on $(0,1)$. In addition, its distribution will usually still depend on θ. However, the distribution of $\widehat{U}$ should approximate that of $U(0,1)$.
The above method for bootstrap confidence sets is nothing but Hall's percentile method, provided we take as root the *location root*

$$R(\mathbf{X}, \psi) = \widehat{\psi} - \psi = \widehat{\psi}(\mathbf{X}) - \psi \ .$$

Thus the above bootstrap confidence sets based on roots represent an extension of Hall's percentile method to other than location roots.

5.1.4 The Iteration or Prepivoting Principle

Beran's double bootstrap or prepivoting idea consists in treating

$$R_1(\mathbf{X}, \psi) = \widehat{U} = F_{\widehat{\theta}}(R(\mathbf{X}, \psi))$$

as another root and in applying the above bootstrap confidence set process with $R_1(\mathbf{X}, \psi)$ as root. Note that $R_1(\mathbf{X}, \psi)$ depends on $\mathbf{X}$ in two ways, once through $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ in $F_{\widehat{\theta}}$ and once through $\mathbf{X}$ in $R(\mathbf{X}, \psi)$. We denote the distribution function of R_1 by $F_{1\theta}$. It is worthwhile to point out again the double dependence of $F_{1\theta}$ on θ, namely through P_θ and $\psi(\theta)$ in

$$F_{1\theta}(x) = P_\theta \left(R_1(\mathbf{X}, \psi(\theta)) \leq x \right) \ .$$

The formal bootstrap procedure consists in estimating $F_{1\theta}(x)$ by $F_{1\widehat{\theta}}(x)$, i.e., by replacing θ with $\widehat{\theta}$. When the functional form of $F_{1\theta}$ is not known one resorts again to simulation as will be explained later.
Denoting the $(1-\alpha)$-quantile of $F_{1\widehat{\theta}}(x)$ by

$$r_{1,1-\alpha}(\widehat{\theta}) = F^{-1}_{1\widehat{\theta}}(1-\alpha)$$

we obtain the following confidence set for ψ

$$C_{1B}(\mathbf{X}, 1-\alpha) = \left\{ \psi : R_1(\mathbf{X}, \psi) \leq r_{1,1-\alpha}(\widehat{\theta}) \right\} = \left\{ \psi : F_{1\widehat{\theta}}(R_1(\mathbf{X}, \psi)) \leq 1-\alpha \right\} \ ,$$

with nominal confidence level $1-\alpha$. Again, the second form of the confidence set $C_{1B}(\mathbf{X}, 1-\alpha)$ shows the appeal to the estimated probability integral transform, since $F_{1\theta}(R_1(\mathbf{X}, \psi))$ is exactly $U(0,1)$, provided $F_{1\theta}$ is continuous. Actually we are dealing with a repeated estimated probability integral transform since R_1 already represented such a transform. It is hoped that this repeated transform

$$R_2(\mathbf{X}, \psi) = F_{1\widehat{\theta}}(R_1(\mathbf{X}, \psi)) = F_{1\widehat{\theta}}\left(F_{\widehat{\theta}}(R(\mathbf{X}, \psi))\right)$$

provides a closer approximation to the $U(0,1)$ distribution than the original single transform

$$R_1(\mathbf{X}, \psi) = F_{\widehat{\theta}}(R(\mathbf{X}, \psi)) \ .$$

Beran refers to the step of going from $R(\mathbf{X}, \psi)$ to $R_1(\mathbf{X}, \psi)$ as prepivoting. Of course, the above process, $R(\mathbf{X}, \psi) \rightarrow R_1(\mathbf{X}, \psi) \rightarrow R_2(\mathbf{X}, \psi)$, can in principle be continued, but this is not very useful in practice, especially when nested simulations are needed to carry out the iteration steps.

5.1.5 Calibrated Confidence Coefficients

The following third form of $C_{1B}(\mathbf{X}, 1-\alpha)$ will not only be more useful in the construction of $C_{1B}(\mathbf{X}, 1-\alpha)$ via the bootstrap simulation, but also in elaborating the connection to Loh's calibration scheme, namely

$$\begin{aligned} C_{1B}(\mathbf{X}, 1-\alpha) &= \left\{\psi : F_{1\widehat{\theta}}(R_1(\mathbf{X}, \psi)) \leq 1-\alpha\right\} \\ &= \left\{\psi : F_{1\widehat{\theta}}\left(F_{\widehat{\theta}}(R(\mathbf{X}, \psi))\right) \leq 1-\alpha\right\} \\ &= \left\{\psi : R(\mathbf{X}, \psi) \leq F^{-1}_{\widehat{\theta}}\left(F^{-1}_{1\widehat{\theta}}(1-\alpha)\right)\right\} \ , \end{aligned}$$

i.e., we compare the original root $R(\mathbf{X}, \psi)$ against an adjusted or recalibrated quantile. This recalibration idea was introduced independently by Loh (1987) and was shown to be equivalent to Beran's prepivoting by DiCiccio and Romano (1988) in the case, where the uncalibrated intervals are the ordinary bootstrap intervals $C_B(\mathbf{X}, 1-\alpha)$. To see this, note that the exact coverage of C_B is

$$P_\theta\left(F_{\widehat{\theta}}(R(\mathbf{X}, \psi)) \leq 1-\alpha\right) = P_\theta\left(R_1(\mathbf{X}, \psi) \leq 1-\alpha\right) = F_{1\theta}\left(1-\alpha\right) \ .$$

By replacing θ by $\widehat{\theta}$ in $F_{1\theta}$ we are invoking the bootstrap principle and get an estimated exact coverage of C_B as $F_{1\widehat{\theta}}(1-\alpha)$. The calibration idea is

to choose the original nominal confidence level in the definition of C_B, now denoted by $1-\alpha_1$, such that the estimated exact coverage of $C_B(\mathbf{X}, 1-\alpha_1)$ becomes $1-\alpha$, i.e.,

$$F_{1\widehat{\theta}}(1-\alpha_1) = 1-\alpha \quad \text{or} \quad 1-\alpha_1 = F^{-1}_{1\widehat{\theta}}(1-\alpha) \; .$$

Thus the recalibrated bootstrap confidence set becomes

$$\begin{aligned} C_B(\mathbf{X}, 1-\alpha_1) &= \left\{\psi : F_{\widehat{\theta}}(R(\mathbf{X}, \psi)) \leq 1-\alpha_1\right\} \\ &= \left\{\psi : F_{\widehat{\theta}}(R(\mathbf{X}, \psi)) \leq F^{-1}_{1\widehat{\theta}}(1-\alpha)\right\} \\ &= \left\{\psi : R(\mathbf{X}, \psi) \leq F^{-1}_{\widehat{\theta}}\left(F^{-1}_{1\widehat{\theta}}(1-\alpha)\right)\right\} \\ &= C_{1B}(\mathbf{X}, 1-\alpha) \; , \end{aligned}$$

which establishes the equivalence.

5.1.6 An Analytical Example

Before going into the simulation aspects of the just described confidence sets we will illustrate the method with Example 2, where one can track analytically what happens. Here we are interested in confidence bounds for

$$\psi = \psi(\theta) = \psi(\mu, \sigma) = \mu$$

and as estimates for μ and σ we consider the maximum likelihood estimates $\widehat{\mu}$ and $\widehat{\sigma}$. As root we consider the location root

$$R(\mathbf{X}, \mu) = \widehat{\mu} - \mu = \bar{X} - \mu \; .$$

Analytically F_θ is found to be

$$F_\theta(x) = P_\theta\left(\bar{X} - \mu \leq x\right) = \Phi\left(\frac{\sqrt{n}x}{\sigma}\right) \; .$$

This leads to

$$R_1(\mathbf{X}, \mu) = F_{\widehat{\theta}}(R(\mathbf{X}, \mu)) = \Phi\left(\frac{\sqrt{n}R(\mathbf{X}, \mu)}{\widehat{\sigma}}\right) = \Phi\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\widehat{\sigma}}\right) \; .$$

Since

$$T_{n-1} = \frac{\sqrt{n}(\bar{X} - \mu)}{\widehat{\sigma}\sqrt{n/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim G_{n-1} \; ,$$

where G_{n-1} represents the Student-t distribution function with $n-1$ degrees of freedom, we find that

$$\begin{aligned} F_{1\theta}(x) = P_\theta\left(R_1(\mathbf{X},\mu) \le x\right) &= P\left(\Phi\left(\sqrt{n/(n-1)}\, T_{n-1}\right) \le x\right) \\ &= G_{n-1}\left(\sqrt{(n-1)/n}\, \Phi^{-1}(x)\right) . \end{aligned}$$

In this specific case F_θ still depends on θ, namely on σ, but $F_{1\theta}$ is independent of θ. Thus

$$F_{1\theta}(x) = F_{1\widehat{\theta}}(x) = F_1(x)$$

and its $(1-\alpha)$-quantile is

$$r_{1,1-\alpha} = r_{1,1-\alpha}(\widehat{\theta}) = \Phi\left(\sqrt{\frac{n}{n-1}}\, t_{n-1}(1-\alpha)\right) ,$$

where $t_{n-1}(1-\alpha) = G_{n-1}^{-1}(1-\alpha)$. The confidence set $C_{1B}(\mathbf{X}, 1-\alpha)$ can now be derived as

$$\begin{aligned} C_{1B}(\mathbf{X}, 1-\alpha) &= \left\{\mu : R_1(\mathbf{X},\mu) \le \Phi\left(\sqrt{\frac{n}{n-1}}\, t_{n-1}(1-\alpha)\right)\right\} \\ &= \left\{\mu : \Phi\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\widehat{\sigma}}\right) \le \Phi\left(\sqrt{\frac{n}{n-1}}\, t_{n-1}(1-\alpha)\right)\right\} \\ &= \left\{\mu : \sqrt{n}\, \frac{\bar{X}-\mu}{S} \le t_{n-1}(1-\alpha)\right\} \\ &= \left\{\mu : \bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha) \le \mu\right\} \\ &= \left[\bar{X} - \frac{S}{\sqrt{n}}\, t_{n-1}(1-\alpha), \infty\right) , \end{aligned}$$

leading to the classical lower confidence bound for μ, with exact coverage $1-\alpha$. Of course, the above derivation appears rather convoluted in view of the usual straightforward derivation of the classical bounds. This convoluted process is not an intrinsic part of the prepivoting method and results only from the analytical tracking of the prepivoting method. When prepivoting is done by simulation, see Section 5.1.7, the derivation of the confidence bounds is conceptually more straightforward, and all the work is in the simulation effort.

A similar convoluted exercise, still in the context of Example 2 and using the prepivot method with the location root $R(\mathbf{X}, \sigma^2) = \widehat{\sigma}^2 - \sigma^2$, leads to the lower bound $n\widehat{\sigma}^2/\chi^2_{n-1}(1-\alpha)$. This coincides with the classical lower confidence bound for σ^2, with exact coverage $1-\alpha$. Here $\chi^2_{n-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the chi-square distribution with $n-1$ degrees of freedom. Here matters would have been even better had we used the scale pivot $R(\mathbf{X}, \sigma^2) = \widehat{\sigma}^2/\sigma^2$ instead. In that case the simple bootstrap confidence set $C_B(\mathbf{X}, 1-\alpha)$ would immediately lead to the classical bounds and bootstrap iteration would not be necessary. This particular example shows that the choice of root definitely improves matters.

It turns out that the above examples can be generalized and in doing so the demonstration of the exact coverage property becomes greatly simplified. However, the derivation of the confidence bounds themselves may still be complicated.

The exact coverage in both the above examples is just a special case of the following general result. In our generic setup let us further assume that

$$R_1(\mathbf{X}, \psi) = F_{\widehat{\theta}}(R(\mathbf{X}, \psi))$$

is an exact pivot with continuous distribution function F_1, which is independent of θ. This pivot assumption is satisfied in both our previous normal examples and it is the reason behind the exact coverage there as well as in this general case. Namely,

$$C_{1B}(\mathbf{X}, 1-\alpha) = \left\{\psi : F_1\left(F_{\widehat{\theta}}(R(\mathbf{X}, \psi))\right) \leq 1-\alpha\right\}$$

has exact coverage since

$$U = F_1\left(F_{\widehat{\theta}}(R(\mathbf{X}, \psi))\right)$$

has the $U(0,1)$ distribution.

5.1.7 Prepivoting by Simulation

When analytical methods fail in determining $F_{1\widehat{\theta}}$ or $F_{\widehat{\theta}}$, we can still do so by bootstrap simulation methods. This was already illustrated in Section 5.1.3 for $F_{\widehat{\theta}}$, but for $F_{1\widehat{\theta}}$ a nested simulation is needed.

In order to approximate $F_{1\widehat{\theta}}(x)$ we will simulate $\mathbf{X}^\star_1, \ldots, \mathbf{X}^\star_B$ from $P_{\widehat{\theta}}$ and compute a bootstrap sample of roots

$$R_1\left(\mathbf{X}^\star_1, \psi(\widehat{\theta})\right), \ldots, R_1\left(\mathbf{X}^\star_B, \psi(\widehat{\theta})\right) ,$$

where we postpone for the moment the discussion of how to compute each such root. Note that $\widehat{\theta}$ has taken the place of θ in $\psi(\widehat{\theta})$ and in $P_{\widehat{\theta}}$, which generated the bootstrap sample.
By the LLN we have that

$$\frac{1}{B}\sum_{i=1}^{B} I_{[R_1\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)\leq x]} \longrightarrow F_{1\widehat{\theta}}(x) \text{ as } B\to\infty .$$

As for the computation of each $R_1\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)$, we will need to employ a second level of bootstrap sampling. Recall that

$$R_1\left(\mathbf{X},\psi(\theta)\right) = F_{\widehat{\theta}}\left(R\left(\mathbf{X},\psi(\theta)\right)\right)$$

and thus

$$R_1\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right) = F_{\widehat{\theta}_i^\star}\left(R\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)\right) ,$$

where $\widehat{\theta}_i^\star = \widehat{\theta}(\mathbf{X}_i^\star)$, with $\mathbf{X}_i^\star$ generated by $P_{\widehat{\theta}}$.
For any $\widehat{\theta}_i^\star$ generate a second level bootstrap sample

$$\mathbf{X}_{ij}^{\star\star}, \quad j=1,\ldots,A$$

and compute the corresponding bootstrap sample of original roots

$$R\left(\mathbf{X}_{ij}^{\star\star},\psi\left(\widehat{\theta}_i^\star\right)\right), \quad j=1,\ldots,A .$$

By the LLN we have that

$$\frac{1}{A}\sum_{j=1}^{A} I_{[R\left(\mathbf{X}_{ij}^{\star\star},\psi\left(\widehat{\theta}_i^\star\right)\right)\leq x]} \longrightarrow F_{\widehat{\theta}_i^\star}(x) \quad \text{as} \quad A\to\infty$$

and thus

$$\begin{aligned}\widehat{R}_{1i} &= \frac{1}{A}\sum_{j=1}^{A} I_{[R\left(\mathbf{X}_{ij}^{\star\star},\psi\left(\widehat{\theta}_i^\star\right)\right)\leq R\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)]} \\ &\longrightarrow F_{\widehat{\theta}_i^\star}\left(R\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)\right) = R_1\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right) \quad \text{as } A\to\infty .\end{aligned}$$

Thus, for large A we can consider $\widehat{R}_{1i}$ as a good approximation to $R_1\left(\mathbf{X}_i^\star,\psi(\widehat{\theta})\right)$.
In the same vein we can, for large B and A, consider

$$\widehat{R}_2(x) = \frac{1}{B}\sum_{i=1}^{B} I_{[\widehat{R}_{1i}\leq x]} = \frac{1}{B}\sum_{i=1}^{B} I_{[R_1(\mathbf{X}_i^\star,\psi(\widehat{\theta})\leq x]}$$

as a good approximation of $F_{1\widehat{\theta}}(x)$. In particular, by sorting the $\widehat{R}_{1i}$ we can obtain their $(1-\alpha)$-quantile $\widehat{\gamma} = \widehat{r}_1(1-\alpha)$ by the usual method and treat it as a good approximation for $\gamma = F^{-1}_{1\widehat{\theta}}(1-\alpha)$. Sorting the first level bootstrap sample

$$R(\mathbf{X}_1^\star, \widehat{\psi}), \ldots, R(\mathbf{X}_B^\star, \widehat{\psi})$$

we can find their $\widehat{\gamma}$-quantile $\widehat{r}_{\widehat{\gamma}}$ as a good approximation to

$$r_\gamma(\widehat{\theta}) = F^{-1}_{\widehat{\theta}}(\gamma) = F^{-1}_{\widehat{\theta}}\left(F^{-1}_{1\widehat{\theta}}(1-\alpha)\right)$$

and we may then consider

$$\widehat{C}_{1B}(\mathbf{X}) = \left\{\psi : R(\mathbf{X}, \psi) \leq \widehat{r}_{\widehat{\gamma}}\right\}$$

as a good approximation to $C_{1B}(\mathbf{X}, 1-\alpha)$ for large A and B. In fact, under some general regularity conditions, legitimizing the double limit $A \to \infty$ and $B \to \infty$, one has

$$\widehat{C}_{1B}(\mathbf{X}) \longrightarrow C_{1B}(\mathbf{X}, 1-\alpha) \quad \text{as} \quad A \to \infty,\ B \to \infty\ .$$

5.1.8 Concluding Remarks

The examples in Section 5.1.6 show that the coverage properties of the prepivoting bootstrap method improved over that of Hall's percentile method. In fact, in these particular examples we wound up with exact coverage probabilities. This exactness is special to these examples and to the generalization given in Section 5.1.6. However, Beran (1987) shows under fairly broad conditions (namely $\mathbf{X} = (X_1, \ldots, X_n)$ being a random sample, the root $R(\mathbf{X}, \psi) = \sqrt{n}(\widehat{\psi} - \psi)$ being asymptotically normal $N(0, \sigma^2(\theta))$, and some more regularity conditions) that the coverage error of $C_{1B}(\mathbf{X}, 1-\alpha)$ is of order $1/n$. Further, the coverage error of $C_{1B}(\mathbf{X}, 1-\alpha)$, when using the Studentized root $\sqrt{n}(\widehat{\psi} - \psi)/\sigma(\widehat{\theta})$, is of order $1/n^{3/2}$. Thus the bootstrap iteration in the prepivot method definitely improves matters.

Just as Hall's percentile method generally is not transformation equivariant, one cannot generally expect this property to hold for the prepivoting method either.

Aside from the simulation and computational burden another minor drawback of the prepivoting method is that nothing is said about the choice of the root. Sometimes roots, like the location root $\widehat{\psi} - \psi$ or the scale root $\widehat{\psi}/\psi$, are quite natural but at other times that is not the case. For example,

when $\psi = \rho$ is the bivariate normal correlation in Example 4, one can hardly treat ρ as location or scale parameter and either of the above two roots is inappropriate. There is of course a natural pivot for ρ, but it is very complicated and difficult to compute. The next section presents a modification of the double bootstrap method which gets around the need of choosing a root by automatically generating a canonical root as part of the process.

5.2 The Automatic Double Bootstrap

This section presents a modification of the double bootstrap method. This method, due to Scholz (1992), gets around the need for choosing a root, by automatically generating a canonical root as part of the process. It is not necessary to know the form of the root. If this canonical root is indeed an exact pivot, this modified double bootstrap method will yield exact coverage confidence bounds. We will introduce the method first in the context of tame pivots and then extend it to general exact pivots and show that the resulting confidence bounds are transformation equivariant. We examine the prepivoting connection and then present a case study that examines the sensitivity of the method to the choice of starting estimates. Finally we discuss the case when exact pivots do not exist and we suggest that the automatic double bootstrap procedure should still work well in an approximate sense. This is illustrated with the classical Behrens-Fisher problem.

5.2.1 Exact Confidence Bounds for Tame Pivots

Suppose $\mathbf{X} \sim P_\theta$ and we are interested in confidence bounds for the real valued functional $\psi = \psi(\theta)$. We also have available the estimate $\widehat{\theta}$ of θ and estimate ψ by $\widehat{\psi} = \psi(\widehat{\theta})$. Throughout this subsection it is assumed that we deal with a special type of pivot R, namely a "tame pivot," a name suggested to me by Antonio Possolo. A tame pivot is a function R of $\widehat{\psi}$ and ψ only, with $R(\widehat{\psi}, \psi)$ having distribution function F independent of θ. Further, we assume this pivot function R to have the following monotonicity properties:

(i) $R(\widehat{\psi}, \psi) \searrow$ in ψ for fixed $\widehat{\psi}$

(ii) $R(\widehat{\psi}, \psi) \nearrow$ in $\widehat{\psi}$ for fixed ψ.

Note that these assumptions do not preclude the presence of nuisance parameters. However, the role of such nuisance parameters is masked in that they

neither appear in the pivot nor influence its distribution F. As such, these parameters are not really a nuisance. The following two examples satisfy the above assumptions and in both cases nuisance parameters are present in the model.

In the first example we revisit Example 4. Here we are interested in confidence bounds for the correlation coefficient $\psi = \psi(\theta) = \rho$. Fortuitously, the distribution function $H_\rho(r)$ of the maximum likelihood estimate $\widehat{\rho}$ is continuous, depends only on the parameter ρ, and is monotone decreasing in ρ for fixed r (see Lehmann 1986, p.340). Further,

$$R(\widehat{\rho}, \rho) = H_\rho(\widehat{\rho}) \sim U(0,1)$$

is a pivot. Thus **(i)** and **(ii)** are satisfied. This example has been examined extensively in the literature and Hall (1992) calls it the "smoking gun" of bootstrap methods, i.e., any good bootstrap method better perform reasonably well on this example. For example, the percentile-t method fails spectacularly here, mainly because Studentizing does not pivot in this case. This question was raised by Reid (1981) in the discussion of Efron (1981).

In the second example we revisit Example 2. Here we are interested in confidence bounds on $\psi = \psi(\theta) = \sigma^2$. Using again maximum likelihood estimates we have that

$$R(\widehat{\psi}, \psi) = \frac{\widehat{\psi}}{\psi} = \frac{\widehat{\sigma}^2}{\sigma^2}$$

is a pivot and satisfies **(i)** and **(ii)**.

If we know the pivot distribution function F and the functional form of R, we can construct exact confidence bounds for ψ as follows. From

$$P_\theta \left(R(\widehat{\psi}, \psi) \leq F^{-1}(1-\alpha) \right) = 1 - \alpha$$

we obtain via monotonicity property **(i)**

$$P_\theta \left(\psi \geq R^{-1}_{\widehat{\psi}} \left(F^{-1}(1-\alpha) \right) \right) = 1 - \alpha \ ,$$

where $\widehat{\psi}_L = R^{-1}_{\widehat{\psi}} \left(F^{-1}(1-\alpha) \right)$ solves

$$R(\widehat{\psi}, \psi) = F^{-1}(1-\alpha) \quad \text{or} \quad F\left(R(\widehat{\psi}, \psi) \right) = 1 - \alpha$$

for ψ. Hence we have in $\widehat{\psi}_L$ an exact $100(1-\alpha)\%$ lower confidence bound for ψ. The dependence of $\widehat{\psi}_L$ on F and R is apparent.

It turns out that it is possible in principle to get the same exact confidence bound without knowing F or R, as long as they exist. This is done at the expense of performing the double bootstrap. Here exactness holds provided both bootstrap simulation sample sizes tend to infinity.
The procedure is as follows. First obtain a bootstrap sample of estimates $\widehat{\psi}_1^\star, \ldots, \widehat{\psi}_B^\star$ by the usual process from $P_{\widehat{\theta}}$. By the LLN we have

$$\widehat{G}_B\left(y|\widehat{\theta}\right) = \frac{1}{B}\sum_{i=1}^{B} I_{[\widehat{\psi}_i^\star \leq y]} \longrightarrow P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq y\right) \quad \text{as} \quad B \to \infty \, .$$

Using this empirical distribution function $\widehat{G}_B\left(y|\widehat{\theta}\right)$ we are able to approximate $P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq y\right)$ to any accuracy by just taking B large enough. With the understanding of this approximation we will thus use $\widehat{G}_B\left(y|\widehat{\theta}\right)$ and $P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq y\right)$ interchangeably.
From monotonicity property **(ii)** we then have

$$P_{\widehat{\theta}}\left(\widehat{\psi}^\star \leq y\right) = P_{\widehat{\theta}}\left(R(\widehat{\psi}^\star, \widehat{\psi}) \leq R(y, \widehat{\psi})\right) = F\left(R(y, \widehat{\psi})\right) \, .$$

Next, given a value $\widehat{\theta}_i^\star$ and $\widehat{\psi}_i^\star = \psi(\widehat{\theta}_i^\star)$, we obtain a second level bootstrap sample of estimates

$$\widehat{\psi}_{i1}^{\star\star}, \ldots, \widehat{\psi}_{iA}^{\star\star} \quad \text{from} \quad P_{\widehat{\theta}_i^\star} \, .$$

Exploiting again the monotonicity property **(ii)** we have

$$P_{\widehat{\theta}_i^\star}\left(\widehat{\psi}_i^{\star\star} \leq y\right) = P_{\widehat{\theta}_i^\star}\left(R(\widehat{\psi}_i^{\star\star}, \widehat{\psi}_i^\star) \leq R(y, \widehat{\psi}_i^\star)\right) = F\left(R(y, \widehat{\psi}_i^\star)\right) \, ,$$

which, for large A, can be approximated by the empirical distribution function of $\widehat{\psi}_{i1}^{\star\star}, \ldots, \widehat{\psi}_{iA}^{\star\star}$, i.e., by

$$\widehat{G}_{1A}\left(y|\widehat{\theta}_i^\star\right) = \frac{1}{A}\sum_{j=1}^{A} I_{[\widehat{\psi}_{ij}^{\star\star} \leq y]} \, .$$

By the LLN this converges to $P_{\widehat{\theta}_i^\star}\left(\widehat{\psi}_i^{\star\star} \leq y\right)$, as $A \to \infty$. Thus, by taking A sufficiently large and using $y = \widehat{\psi}$, we can simulate

$$\widehat{G}_{1A}\left(\widehat{\psi}|\widehat{\theta}_1^\star\right), \ldots, \widehat{G}_{1A}\left(\widehat{\psi}|\widehat{\theta}_B^\star\right)$$

and regard them as equivalent proxy for

$$F\left(R(\widehat{\psi},\widehat{\psi}_1^\star)\right),\ldots,F\left(R(\widehat{\psi},\widehat{\psi}_B^\star)\right)\ .$$

Sorting these values we find the $(1-\alpha)$-quantile by the usual process. The corresponding $\widehat{\psi}^\star=\widehat{\psi}_i^\star=\widehat{\psi}_L^\star$ approximately solves

$$F\left(R(\widehat{\psi},\widehat{\psi}^\star)\right)\approx 1-\alpha\ .$$

This value $\widehat{\psi}_L^\star$ is approximately the same as our previous $\widehat{\psi}_L$, provided A and B are sufficiently large.
The above procedure can be reduced to the following: Find that value $\widehat{\theta}^\star$ and $\widehat{\psi}^\star=\psi(\widehat{\theta}^\star)$, for which

$$P_{\widehat{\theta}^\star}\left(\widehat{\psi}^{\star\star}\le\widehat{\psi}\right)=F\left(R(\widehat{\psi},\widehat{\psi}^\star)\right)\approx 1-\alpha\ .$$

Start with a value $\widehat{\psi}_1^\star$, say $\widehat{\psi}_1^\star=\widehat{\psi}$, and take a convenient value $\widehat{\theta}_1^\star\in\psi^{-1}(\widehat{\psi}_1^\star)$ so that $\psi(\widehat{\theta}_1^\star)=\widehat{\psi}_1^\star$. Using a second level bootstrap sample of size A, evaluate or approximate

$$F_1=F\left(R(\widehat{\psi},\widehat{\psi}_1^\star)\right)=P_{\widehat{\theta}_1^\star}\left(\widehat{\psi}^{\star\star}\le\widehat{\psi}\right)\quad\text{by}\quad\widehat{G}_{1A}\left(\widehat{\psi}|\widehat{\theta}_1^\star\right)\ .$$

This is then iterated by trying new values of $\widehat{\psi}^\star$, i.e., $\widehat{\psi}_1^\star$, $\widehat{\psi}_2^\star$, ... Since

$$F\left(R(\widehat{\psi},\widehat{\psi}^\star)\right)\searrow\quad\text{in}\quad\widehat{\psi}^\star$$

one should be able to employ efficient root finding algorithms for solving

$$F\left(R(\widehat{\psi},\widehat{\psi}^\star)\right)=1-\alpha\ ,$$

i.e., use far fewer than the originally indicated AB bootstrap iterations. It seems reasonable that kA iterations will be sufficient with $A\approx 1000$ and $k\approx 10$ to 20.
Note that in this procedure we only need to evaluate $\widehat{G}_{1A}\left(\widehat{\psi}|\widehat{\theta}_i^\star\right)$, which in turn only requires that we know how to evaluate the estimates $\widehat{\psi}$, $\widehat{\psi}^\star$, or $\widehat{\psi}^{\star\star}$. No knowledge of the pivot function R or its distribution function F is required. For the previously discussed bivariate normal correlation example, there exists a tame pivot. Therefore we either can, through massive simulation of computationally simple calculations of $\widehat{\rho}$, obtain the exact confidence

bound through the above bootstrap process, or in its place use the computationally difficult analytical process of evaluating the distribution function $H_\rho(x)$ of $\widehat{\rho}$ and solving

$$H_\rho\left(\widehat{\rho}\right) = \alpha$$

for $\rho = \widehat{\rho}_U$. Then

$$P_\rho\left(\rho \leq \widehat{\rho}_U\right) = P_\rho\left(H_\rho\left(\widehat{\rho}\right) \geq \alpha\right) = 1 - \alpha\ .$$

5.2.2 The General Pivot Case

For the general pivot case we assume that the indexing parameter can be reparametrized in the form $\theta = (\psi, \eta)$, i.e., the quantity ψ of interest is just a particular real valued component of θ and the remainder η of θ acts as a vector of nuisance parameters. Again we have estimates $\widehat{\theta} = (\widehat{\psi}, \widehat{\eta})$ and we denote the distribution function of $\widehat{\psi}$ by

$$D_{\psi,\eta}(y) = P_\theta\left(\widehat{\psi} \leq y\right)\ .$$

Motivated by the probability integral transform result, $D_{\psi,\eta}(\widehat{\psi}) \sim U(0,1)$ for continuous $D_{\psi,\eta}$, we make the following general pivot assumption:

(*V*) $D_{\psi,\widehat{\eta}}\left(\widehat{\psi}\right)$ is a pivot, i.e., has a distribution function H which does not depend on unknown parameters, and $D_{\psi,\widehat{\eta}}\left(\widehat{\psi}\right) \searrow$ in ψ for fixed $\widehat{\psi}$ and $\widehat{\eta}$.

The tame pivot case examined in the previous section satisfies (*V*) if F is continuous. Namely,

$$D_\theta(y) = P_\theta\left(\widehat{\psi} \leq y\right) = P_\theta\left(R(\widehat{\psi}, \psi) \leq R(y, \psi)\right) = F(R(y, \psi))$$

and

$$D_{\psi,\widehat{\eta}}(\widehat{\psi}) = F(R(\widehat{\psi}, \psi)) \sim U(0,1)$$

is a pivot and is decreasing in ψ. Here $\widehat{\eta}$ does not affect $D_{\psi,\widehat{\eta}}(\widehat{\psi})$.
Example 2 of a normal random sample illustrates the situation of estimated nuisance parameters. Here we are interested in confidence bounds for the p-quantile $\psi = \psi(\theta) = \mu + z_p\sigma$, where z_p is the standard normal p-quantile.

We can think of θ as reparametrized in terms of ψ and σ and again we use maximum likelihood estimates $\widehat{\psi}$ and $\widehat{\sigma}$ for ψ and σ. We have that

$$\frac{\widehat{\psi}-\psi}{\widehat{\sigma}} \quad \text{and} \quad \frac{\widehat{\psi}-\psi}{\sigma}$$

are both pivots with respective c.d.f.'s G_1 and G_2 and

$$D_\theta(y) = P_\theta\left(\widehat{\psi} \leq y\right) = G_2\left(\frac{y-\psi}{\sigma}\right) \ .$$

Thus

$$D_{\psi,\widehat{\sigma}}(\widehat{\psi}) = G_2\left(\frac{\widehat{\psi}-\psi}{\widehat{\sigma}}\right) \sim G_2\left(G_1^{-1}(U)\right) \ ,$$

where $U \sim U(0,1)$.
This example generalizes easily. Assume that there is a function $R(\widehat{\psi},\psi,\eta)$ which is a pivot, i.e., has distribution function G_2, and is decreasing in ψ and increasing in $\widehat{\psi}$. Suppose further that $R(\widehat{\psi},\psi,\widehat{\eta})$ is also a pivot with distribution function G_1. Then again our general pivot assumption (V) is satisfied. This follows from

$$D_{\psi,\eta}(y) = P_{\psi,\eta}\left(\widehat{\psi} \leq y\right) = P_{\psi,\eta}\left(R(\widehat{\psi},\psi,\eta) \leq R(y,\psi,\eta)\right) = G_2\left(R(y,\psi,\eta)\right)$$

and thus

$$D_{\psi,\widehat{\eta}}\left(\widehat{\psi}\right) = G_2\left(R\left(\widehat{\psi},\psi,\widehat{\eta}\right)\right) = G_2\left(G_1^{-1}(U)\right)$$

is a pivot which is decreasing in ψ.
Given the general pivot assumption (V), it is possible to construct exact lower confidence bounds for ψ as follows:

$$1-\alpha = P_\theta\left(D_{\psi,\widehat{\eta}}(\widehat{\psi}) \leq H^{-1}(1-\alpha)\right) = P_\theta\left(\psi \geq \widehat{\psi}_L\right) \ ,$$

where $\widehat{\psi}_L$ solves

$$D_{\psi,\widehat{\eta}}(\widehat{\psi}) = H^{-1}(1-\alpha) \quad \text{or} \quad H\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right) = 1-\alpha \tag{4}$$

for $\psi = \widehat{\psi}_L$. This, however, requires knowledge of both H and D.
We observe that $\widehat{\psi}_L$ is transformation equivariant. To show this, let g be a strictly increasing transform of ψ into $\tau = g(\psi)$. We use $\widehat{\tau} = g(\widehat{\psi})$ as our

estimate of τ. Then the above procedure applied to $\widehat{\tau}$ and with $\theta = (\psi, \eta)$ reparametrized to $\vartheta = (\tau, \eta)$ yields $\widehat{\tau}_L = g(\widehat{\psi}_L)$.
This is seen as follows. Denote the reparametrized probability model by $\widetilde{P}_{\tau,\eta}$, which is equivalent to $P_{g^{-1}(\tau),\eta}$. The distribution function of $\widehat{\tau}$ is

$$\begin{aligned}\widetilde{D}_{\tau,\eta}(y) &= \widetilde{P}_{\tau,\eta}\left(\widehat{\tau} \le y\right) = \widetilde{P}_{\tau,\eta}\left(g\left(\widehat{\psi}\right) \le y\right) = \widetilde{P}_{\tau,\eta}\left(\widehat{\psi} \le g^{-1}(y)\right) \\ &= P_{g^{-1}(\tau),\eta}\left(\widehat{\psi} \le g^{-1}(y)\right) = D_{g^{-1}(\tau),\eta}\left(g^{-1}(y)\right) \ ,\end{aligned}$$

so that

$$\widetilde{D}_{\tau,\widehat{\eta}}(\widehat{\tau}) = D_{g^{-1}(\tau),\widehat{\eta}}\left(g^{-1}\left(g\left(\widehat{\psi}\right)\right)\right) = D_{g^{-1}(\tau),\widehat{\eta}}\left(\widehat{\psi}\right)$$

is an exact pivot with same c.d.f. H as $D_{\psi,\widehat{\eta}}(\widehat{\psi})$. Solving

$$1 - \alpha = H\left(\widetilde{D}_{\tau,\widehat{\eta}}(\widehat{\tau})\right) = H\left(D_{g^{-1}(\tau),\widehat{\eta}}\left(\widehat{\psi}\right)\right)$$

for $\tau = \widehat{\tau}_L$ yields

$$g^{-1}\left(\widehat{\tau}_L\right) = \widehat{\psi}_L \quad \text{or} \quad \widehat{\tau}_L = g(\widehat{\psi}_L) \ ,$$

as was to be shown.
By appropriate double bootstrapping we can achieve the same objective, namely finding $\widehat{\psi}_L$, without knowing H or D. The double bootstrap we employ here is a slight variant of the commonly used one. There are two parts to the procedure. The first part obtains $H^{-1}(1-\alpha)$ to any accuracy for large enough A and B and the second consists of the iterative solution of Equation (4).
We start by generating the first level bootstrap sample $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ from P_{ψ_0,η_0} for some choice of ψ_0 and η_0. Typically, for reasons to be discussed in Section 5.2.5, one would take $(\psi_0, \eta_0) = (\widehat{\psi}, \widehat{\eta})$. However, for now we will stay with the arbitrary starting choice (ψ_0, η_0).
From these bootstrap data sets we obtain the first level bootstrap sample of estimates, i.e.,

$$(\widehat{\psi}_i^\star, \widehat{\eta}_i^\star) \ , \quad i = 1, \ldots, B \ .$$

For each $i = 1, \ldots, B$ obtain a second level bootstrap data sample $\mathbf{X}_{i1}^{\star\star}, \ldots, \mathbf{X}_{iA}^{\star\star}$ from $P_{\psi_0,\widehat{\eta}_i^\star}$ (not from $P_{\widehat{\psi}_i^\star,\widehat{\eta}_i^\star}$ as one might usually do it) and compute the corresponding second level bootstrap sample of estimates

$$(\widehat{\psi}_{ij}^{\star\star}, \widehat{\eta}_{ij}^{\star\star}) \ , \quad j = 1, \ldots, A \ , \quad i = 1, \ldots, B \ .$$

By the LLN, as $A \to \infty$, we have

$$\widehat{D}_{iA} = \frac{1}{A} \sum_{j=1}^{A} I_{[\widehat{\psi}_{ij}^{\star\star} \leq \widehat{\psi}_i^{\star}]} \longrightarrow P_{\psi_0, \widehat{\eta}_i^{\star}} \left(\widehat{\psi}_i^{\star\star} \leq \widehat{\psi}_i^{\star} \right) = D_{\psi_0, \widehat{\eta}_i^{\star}} \left(\widehat{\psi}_i^{\star} \right) \sim H \ .$$

The latter distributional assertion derives from the pivot assumption (V) and from the fact that $(\widehat{\psi}_i^{\star}, \widehat{\eta}_i^{\star})$ arises from P_{ψ_0, η_0}. Again appealing to the LLN we have

$$\frac{1}{B} \sum_{i=1}^{B} I_{[D_{\psi_0, \widehat{\eta}_i^{\star}}(\widehat{\psi}_i^{\star}) \leq y]} \longrightarrow H(y) \ \text{ as } \ B \to \infty \ ,$$

and thus we can consider

$$\frac{1}{B} \sum_{i=1}^{B} I_{[\widehat{D}_{iA} \leq y]}$$

as a good approximation to $H(y)$. From this approximation we can obtain $H^{-1}(1-\alpha)$. This is done by sorting the sample $\widehat{D}_{iA}, i = 1, \ldots, A$ and finding its $(1-\alpha)$-quantile by the usual process.
Now comes the second part of the procedure. For some value ψ_1 (one could start here with $\psi_1 = \widehat{\psi}$) generate

$$\mathbf{X}_1^{\circ}, \ldots, \mathbf{X}_N^{\circ} \ \text{ i.i.d. } \sim P_{\psi_1, \widehat{\eta}}$$

and get the bootstrap sample of resulting estimates

$$\widehat{\psi}_{11}^{\circ}, \ldots, \widehat{\psi}_{1N}^{\circ} \ .$$

By the LLN we have

$$\widehat{D}_N^{\circ}(\psi_1) = \frac{1}{N} \sum_{i=1}^{N} I_{[\widehat{\psi}_{1i}^{\circ} \leq \widehat{\psi}]} \longrightarrow D_{\psi_1, \widehat{\eta}}(\widehat{\psi}) \ .$$

Using the monotonicity of $D_{\psi, \widehat{\eta}}(\widehat{\psi})$ in ψ, a few iterations over $\psi_1, \psi_2, \ldots$ should quickly lead to a solution of the equation

$$D_{\psi, \widehat{\eta}}(\widehat{\psi}) \approx \widehat{D}_N^{\circ}(\psi) = H^{-1}(1-\alpha) \ .$$

For large A, B, N this solution is practically identical with the exact lower confidence bound $\widehat{\psi}_L$. If this latter process takes k iterations we will have performed $AB + kN$ bootstrap iterations. This is by no means efficient and it is hoped that future work will make the computational aspects of this approach more practical.

5.2.3 The Prepivoting Connection

We now examine the connection to Beran's prepivoting approach and ,by equivalence, also to Loh's calibrated confidence sets, see Section 5.1.5, Suppose we have a specified root function $R(\widehat{\psi}, \psi) = R(\widehat{\psi}(\mathbf{X}), \psi)$ with distribution function $F_{\psi,\eta}(x)$. This is somewhat more special than Beran's general root concept $R(\mathbf{X}, \psi)$. Suppose now that the following assumptions hold

$(V^\star)$ $F_{\psi,\widehat{\eta}}(R(\widehat{\psi}, \psi))$ is a pivot,
$R(\widehat{\psi}, \psi) \nearrow$ in $\widehat{\psi}$ for fixed ψ and
$F_{\psi,\widehat{\eta}}(R(\widehat{\psi}, \psi)) \searrow$ in ψ for fixed $\widehat{\psi}$ and $\widehat{\eta}$.

Then $(V^\star)$ implies (V), since

$$D_{\psi,\eta}(x) = P_{\psi,\eta}(\widehat{\psi} \leq x) = P_{\psi,\eta}(R(\widehat{\psi}, \psi) \leq R(x, \psi)) = F_{\psi,\eta}(R(x, \psi))$$

and

$$D_{\psi,\widehat{\eta}}(\widehat{\psi}) = F_{\psi,\widehat{\eta}}(R(\widehat{\psi}, \psi))$$

is a pivot by assumption.
When F does not depend on ψ, i.e., when the root function is successful in eliminating ψ from the distribution of R, then one can replace

$$F_{\psi,\widehat{\eta}}(R(\widehat{\psi}, \psi)) \searrow \text{ in } \psi \text{ for fixed } \widehat{\psi} \text{ and } \widehat{\eta}$$

in $(V^\star)$ by the more natural assumption

$$R(\widehat{\psi}, \psi) \searrow \text{ in } \psi \text{ for fixed } \widehat{\psi}.$$

In contrast to the pivot assumption in $(V^\star)$, Beran's prepivoting idea treats

$$F_{\widehat{\psi},\widehat{\eta}}(R(\widehat{\psi}, \psi))$$

as pivotal or nearly pivotal, its distribution being generated via bootstrapping. The difference in the two approaches consists in how the subscript ψ on F is treated. Often it turns out that F depends only on the subscript η and the above distinction does not manifest itself. In those cases Beran's prepivoting will lead to exact confidence bounds as well, provided $(V^\star)$ holds. For example, in the situation of Example 2 with $\psi = \mu + z_p\sigma$ and

$$R\left(\widehat{\psi}, \psi\right) = \widehat{\psi} - \psi = \widehat{\mu} + z_p\widehat{\sigma} - (\mu + z_p\sigma)$$

as root, Beran's prepivoting will lead to exact confidence bounds, since the distribution of R depends only on the nuisance parameter $\eta = \sigma$.
In contrast, consider Example 4 with $\psi = \rho$. If we take the root $R(\widehat{\rho}, \rho) = \widehat{\rho} - \rho$, then

$$F_\rho(x) = P_\rho(\widehat{\rho} - \rho \leq x) = H_\rho(x + \rho)$$

with H_ρ denoting the c.d.f. of $\widehat{\rho}$. Here the assumption $(V^\star)$ is satisfied since

$$F_\rho(\widehat{\rho} - \rho) = H_\rho(\widehat{\rho})$$

is a pivot. However,

$$F_{\widehat{\rho}}(\widehat{\rho} - \rho) = H_{\widehat{\rho}}(\widehat{\rho} - \rho + \widehat{\rho})$$

appears not to be a pivot, although we have not verified this. This difference is mostly due to the badly chosen root. If we had taken as root $R(\widehat{\rho}, \rho) = H_\rho(\widehat{\rho})$, then the distinction would not arise. In fact, in that case R itself is already a pivot. However, this particular root function is not trivial and that points out the other difference between Beran's prepivoting and the automatic double bootstrap. In the latter method no knowledge of an "appropriate" root function is required.
As a complementary example consider Example 2 with the root $R = \sqrt{n}(s^2 - \sigma^2)$ for the purpose of constructing confidence bounds for σ^2. Let χ_f denote the c.d.f. of a chi-square distribution with f degrees of freedom. Then

$$F_{\mu,\sigma^2}(x) = P_{\mu,\sigma^2}\left(\sqrt{n}(s^2 - \sigma^2) \leq x\right) = \chi_{n-1}\left((n-1)\left(1 + \frac{x}{\sqrt{n}\sigma^2}\right)\right) .$$

Clearly

$$\begin{aligned} F_{\widehat{\mu},\sigma^2}(R) &= \chi_{n-1}\left((n-1)\left(1 + \frac{\sqrt{n}(s^2 - \sigma^2)}{\sqrt{n}\sigma^2}\right)\right) \\ &= \chi_{n-1}\left(\frac{(n-1)s^2}{\sigma^2}\right) \sim U(0,1) \end{aligned}$$

is a pivot which will lead to the classical lower bound for σ^2. On the other hand, the iterated root

$$\begin{aligned} R_{1,n}(\sigma^2) = F_{\widehat{\mu},s^2}(R) &= \chi_{n-1}\left((n-1)\left(1 + \frac{\sqrt{n}(s^2 - \sigma^2)}{\sqrt{n}s^2}\right)\right) \\ &= \chi_{n-1}\left((n-1)\left(2 - \frac{\sigma^2}{s^2}\right)\right) \end{aligned}$$

is a pivot as well, with distribution function

$$F_{1,n}(x) = \chi_{n-1}\left((n-1)\left(2 - \frac{\chi_{n-1}^{-1}(x)}{n-1}\right)^{-1}\right) \quad \text{for } 0 < x \leq \chi_{n-1}(2(n-1))$$

and $F_{1,n}(0) = \chi_{n-1}((n-1)/2)$, $F_{1,n}(x) = 0$ for $x < 0$ and $F_{1,n}(x) = 1$ for $x \geq \chi_{n-1}(2(n-1))$. For $\gamma \geq \chi_{n-1}((n-1)/2)$ the set

$$B_{1,n} = \left\{\sigma^2 : F_{1,n}(R_{1,n}) \leq \gamma\right\} = \left[(n-1)s^2/\chi_{n-1}^{-1}(\gamma), \infty\right)$$

yields the classical lower confidence bound, but for $\gamma < \chi_{n-1}((n-1)/2)$ the set $B_{1,n}$ is empty. This quirk was overlooked in Beran's (1987) treatment of this example. For large n the latter case hardly occurs, unless we deal with small γ's, i.e., with upper confidence bounds.

5.2.4 Sensitivity to Choice of Estimates

In this section we will use Example 2 with $\psi = \mu + z_p\sigma$ to illustrate the sensitivity of the pivot method and thus of the proposed automatic double bootstrap method to the choice of starting estimates.
In this example it is instructive to analyze to what extent the form of the estimate $(\widehat{\psi}, \widehat{\sigma})$ affects the form of the lower bound $\widehat{\psi}_L$ for ψ which results from the pivot method.
It is obvious that the lower bound will indeed be different, if we start out with location and scale estimates, which are different in character from that of the maximum likelihood estimates. For example, as location scale estimates one might use the sample median and range or various other robust alternatives. Here we will analyze the more limited situation where we use as estimates of ψ and σ

$$\widehat{\psi} = \bar{X} + ks \quad \text{and} \quad \widehat{\sigma} = rs = r\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$$

for some known constants k and $r > 0$. In question here is the sensitivity of the resulting automatic double bootstrap lower bound $\widehat{\psi}_L$ with respect to k and r. This issue is similar but not the same as that of transformation equivariance.
It turns out that $\widehat{\psi}_L$ does not depend on k or r, i.e., the result is always the same, namely the classical lower confidence bound for ψ. For example, it does

not matter whether we estimate σ by the m.l.e. or by s. More remarkable is the fact that we could have started with the very biased starting estimate $\widehat{\psi} = \bar{X}$, corresponding to $k = 0$, with the same final lower confidence bound. It is possible that there is a general theorem hidden behind this that would more cleanly dispose of the following convoluted argument for this result. This argument fills the remainder of this section and may be skipped.
Recalling $\psi = \mu + z_p\sigma$, one easily derives

$$\begin{aligned} D_{\psi,\sigma}(x) &= P_{\psi,\sigma}\left(\widehat{\psi} \le x\right) = P_{\psi,\sigma}\left(\bar{X} + ks \le x\right) \\ &= P_{\psi,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} + \frac{\sqrt{n}(\mu - x)}{\sigma} \le -ks\sqrt{n}/\sigma\right) \\ &= G_{n-1,\sqrt{n}(\mu-x)/\sigma}(-k\sqrt{n}) \\ &= G_{n-1,\sqrt{n}(\psi-x)/\sigma - z_p\sqrt{n}}(-k\sqrt{n}) \ , \end{aligned} \tag{5}$$

where $G_{f,\delta}(x)$ denotes the noncentral Student-t distribution with f degrees of freedom and noncentrality parameter δ.
Next note that

$$\begin{aligned} D_{\psi,\widehat{\sigma}}(\widehat{\psi}) &= G_{n-1,\sqrt{n}(\psi-\widehat{\psi})/\widehat{\sigma} - z_p\sqrt{n}}(-k\sqrt{n}) \\ &= G_{n-1,-z_p\sqrt{n}-V/r}(-k\sqrt{n}) \ , \end{aligned}$$

where

$$\begin{aligned} V = \frac{\sqrt{n}(\widehat{\psi}-\psi)}{s} &= \frac{\sqrt{n}(\bar{X}-\mu-z_p\sigma)}{s} + k\sqrt{n} \\ &= k\sqrt{n} + T_{n-1,-z_p\sqrt{n}} \end{aligned}$$

and $T_{f,\delta}$ is a random variable with distribution function $G_{f,\delta}(x)$. The distribution function H of $D_{\psi,\widehat{\sigma}}(\widehat{\psi})$ can be expressed more or less explicitly as

$$H(y) = P\left(D_{\psi,\widehat{\sigma}}(\widehat{\psi}) \le y\right) = P\left(-z_p\sqrt{n} - V/r \ge \delta(n-1, -k\sqrt{n}, y)\right) \ ,$$

where $\delta_y = \delta(n-1, -k\sqrt{n}, y)$ solves

$$G_{n-1,\delta_y}(-k\sqrt{n}) = y \ .$$

Using the above representation for V we have

$$\begin{aligned} H(y) &= P\left(T_{n-1,-z_p\sqrt{n}} \le -rz_p\sqrt{n} - r\delta_y - k\sqrt{n}\right) \\ &= G_{n-1,-z_p\sqrt{n}}\left(-\sqrt{n}(rz_p + k) - r\delta(n-1, -k\sqrt{n}, y)\right) \ . \end{aligned}$$

Solving $H(y_{1-\alpha}) = 1-\alpha$ for $y_{1-\alpha} = H^{-1}(1-\alpha)$ we get

$$t_{n-1,-z_p\sqrt{n},1-\alpha} = -\sqrt{n}(rz_p + k) - r\delta(n-1, -k\sqrt{n}, y_{1-\alpha})$$

or

$$-(\sqrt{n}(rz_p + k) + t_{n-1,-z_p\sqrt{n},1-\alpha})/r = \delta(n-1, -k\sqrt{n}, y_{1-\alpha})$$

where $t_{f,\delta,1-\alpha}$ is the $(1-\alpha)$-quantile of $G_{f,\delta}(x)$. Using the defining equation for δ_y we get

$$H^{-1}(1-\alpha) = G_{n-1,-(\sqrt{n}(rz_p+k)+t_{n-1,-z_p\sqrt{n},1-\alpha})/r}(-k\sqrt{n}) \ .$$

Solving

$$H^{-1}(1-\alpha) = D_{\psi,\widehat{\sigma}}(\widehat{\psi})$$

or

$$G_{n-1,-(\sqrt{n}(rz_p+k)+t_{n-1,-z_p\sqrt{n},1-\alpha})/r}(-k\sqrt{n}) = G_{n-1,\sqrt{n}(\psi-\widehat{\psi})/\widehat{\sigma}-z_p\sqrt{n}}(-k\sqrt{n})$$

for $\psi = \widehat{\psi}_L$ we find

$$(\sqrt{n}(rz_p + k) + t_{n-1,-z_p\sqrt{n},1-\alpha})/r = \sqrt{n}(\widehat{\psi} - \psi)/\widehat{\sigma} + z_p\sqrt{n}$$

or

$$\widehat{\psi}_L = \psi = \widehat{\psi} - ks - \frac{s}{\sqrt{n}}\, t_{n-1,-z_p\sqrt{n},1-\alpha} = \bar{X} - \frac{s}{\sqrt{n}}\, t_{n-1,-z_p\sqrt{n},1-\alpha} \ ,$$

which is the conventional $100(1-\alpha)\%$ lower confidence bound for ψ.

5.2.5 Approximate Pivots and Iteration

Previously it was shown that under the general pivot assumption (V) the automatic double bootstrap closes the loop as far as exact confidence bounds are concerned. It is noteworthy in this double bootstrap procedure that we have complete freedom in choosing (ψ_0, η_0). This freedom arises from the pivot assumption. The pivot assumption is a strong one and usually not satisfied. However, in many practical situations one may be willing to assume that there is an approximate local pivot. By "local" we mean that the statement

$$\text{“}D_{\psi,\widehat{\eta}}(\widehat{\psi}) \quad \textit{is approximately distribution free}\text{”}$$

holds in a neighborhood of the true unknown parameter θ. Since presumably $\widehat{\theta}$ is our best guess at θ, we may as well start our search for $H^{-1}(1-\alpha)$ as close as possible to θ, namely with $\theta_0 = (\psi_0, \eta_0) = \widehat{\theta}$, in order to take greatest advantage of the closeness of the used approximation. To emphasize this we write

$$H_{\widehat{\theta}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right) = 1-\alpha$$

as the equation that needs to be solved for ψ to obtain the $100(1-\alpha)\%$ lower bound $\widehat{\psi}_L$ for ψ. Of course, the left side of this equation will typically no longer have a uniform distribution on $(0,1)$. Following Beran (1987) one could iterate this procedure further. If

$$H_{\widehat{\theta}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right) \sim H_{2,\theta}$$

with $H_{2,\widehat{\theta}}\left(H_{\widehat{\theta}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right)\right)$ hopefully more uniform than $H_{\widehat{\theta}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right)$, one could then try for an adjusted lower bound by solving

$$H_{2,\widehat{\theta}}\left(H_{\widehat{\theta}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right)\right) = 1-\alpha$$

for $\psi = \widehat{\psi}_{2,L}$. This process can be further iterated in obvious fashion, but whether this will be useful in small sample situations is questionable. What would such an iteration converge to in the specific situation to be examined next?

As illustration of the application of our method to an approximate pivot situation we will consider the Behrens-Fisher problem, which was examined by Beran (1988) in a testing context from an asymptotic rate perspective.

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from respective $N(\mu, \sigma_1^2)$ and $N(\nu, \sigma_2^2)$ populations. Of interest are confidence bounds for $\psi = \mu - \nu$. Since we do not assume $\sigma_1 = \sigma_2$ we are faced with the classical Behrens-Fisher problem.

We will examine how the automatic double bootstrap or pivot method attacks this problem. We can reparametrize the above model in terms of (ψ, η), where $\mu = \psi + \nu$ and $\eta = (\nu, \sigma_1, \sigma_2)$. As natural estimate of ψ we take $\widehat{\psi} = \bar{X} - \bar{Y}$ and as estimate for η we take $\widehat{\eta} = (\bar{Y}, s_1, s_2)$, where s_i^2 is the usual unbiased estimate of σ_i^2. The distribution function of $\widehat{\psi}$ is

$$D_{\psi,\eta}(x) = P_{\psi,\eta}\left(\bar{X} - \bar{Y} \leq x\right) = \Phi\left(\frac{x-\psi}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right) .$$

The distribution function H_ρ of

$$D_{\psi,\widehat{\eta}}(\widehat{\psi}) = \Phi\left(\frac{\widehat{\psi}-\psi}{\sqrt{s_1^2/m + s_2^2/n}}\right)$$

depends on the unknown parameters through

$$\rho = \rho(\sigma_1^2, \sigma_2^2) = \frac{n\sigma_1^2}{n\sigma_1^2 + m\sigma_2^2} \; .$$

Thus assumption (V) is violated.
Traditional solutions to the problem involve approximating the distribution function $G_\rho(x) = H_\rho(\Phi(x))$ of

$$T = \frac{\widehat{\psi}-\psi}{\sqrt{s_1^2/m + s_2^2/n}}$$

and in the process replace the unknown ρ by $\widehat{\rho} = \rho(s_1^2, s_2^2)$. This is done for example in Welch's solution (Welch (1947) and Aspin (1949)), where G_ρ is approximated by a Student t-distribution function $F_f(t)$ with $f = f(\rho)$ degrees of freedom with

$$f(\rho) = \left(\frac{\rho^2}{m-1} + \frac{(1-\rho)^2}{n-1}\right)^{-1} \; .$$

As a second approximation step one then replaces the unknown ρ by $\widehat{\rho}$, i.e., one estimates f by $\widehat{f} = f(\widehat{\rho})$. This leads to the following lower confidence bound for ψ:

$$\widehat{\psi}_{WL} = \widehat{\psi} - F_{\widehat{f}}^{-1}(1-\alpha)\sqrt{s_1^2/m + s_2^2/n} \; .$$

Recall that in the first phase of the automatic double bootstrap method we could start the process of finding H_ρ with any (ψ_0, η_0). This would result in H_{ρ_0}. This is reasonable as long as H does not depend on unknown parameters. By taking as starting values $(\psi_0, \eta_0) = (\widehat{\psi}, \widehat{\eta})$ we wind up with a determination of $H_{\widehat{\rho}}$ instead. Thus the character of H is maintained and is not approximated. The only approximation that takes place is that of replacing the unknown ρ by $\widehat{\rho}$. Whether this actually improves the coverage properties over those of the Welch solution remains to be seen. There is of

course the possibility that the two approximation errors in Welch's solution cancel each other out to some extent.
The second phase of the pivot or automatic double bootstrap method stipulates that we solve

$$1-\alpha = H_{\widehat{\rho}}\left(D_{\psi,\widehat{\eta}}(\widehat{\psi})\right) = G_{\widehat{\rho}}\left(\frac{\widehat{\psi}-\psi}{\sqrt{s_1^2/m+s_2^2/n}}\right)$$

for $\psi = \widehat{\psi}_L$, which yields the following $100(1-\alpha)\%$ lower bound for ψ

$$\widehat{\psi}_L = \widehat{\psi} - G^{-1}_{\widehat{\rho}}(1-\alpha)\sqrt{s_1^2/m+s_2^2/n}\ .$$

Beran (1988) arrives at exactly the same bound (although in a testing context) by simple bootstrapping. However, he started out with the Studentized test statistic T, which thus is one step ahead in the game. It is possible to analyze the true coverage probabilities for $\widehat{\psi}_L$ and $\widehat{\psi}_{WL}$, although the evaluation of the analytical formulae for these coverage probabilities requires substantial numerical effort.
These analytical formulae are derived by using a well known conditioning device, see Fleiss (1971) for a recent account of details. The formula for the exact coverage probability for $\widehat{\psi}_L$ is as follows

$$\begin{aligned} K_\rho(1-\alpha) &= P_\rho\left(\widehat{\psi}_L \le \psi\right) \\ &= \int_0^1 b(w) F_g\left(G^{-1}_{\widehat{\rho}(w)}(1-\alpha)\sqrt{ga_1(\rho)w + ga_2(\rho)(1-w)}\right) dw \end{aligned}$$

with

$$g = m+n-2\ , \quad a_1(\rho) = \frac{\rho}{m-1}\ , \quad a_2(\rho) = \frac{1-\rho}{n-1}\ .$$

$$b(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, w^{\alpha-1}(1-w)^{\beta-1} I_{[0,1]}(w)$$

is the beta density with $\alpha = (m-1)/2$ and $\beta = (n-1)/2$ and

$$\widehat{\rho}(w) = \frac{w\rho(n-1)}{w\rho(n-1) + (1-w)(1-\rho)(m-1)}\ .$$

$G^{-1}_\rho(p)$ is the inverse of

$$G_\rho(x) = P_\rho(T \le x) = \int_0^1 b(u) F_g\left(x\sqrt{ga_1(\rho)u + ga_2(\rho)(1-u)}\right) du\ .$$

The corresponding formula for the exact coverage of $\widehat{\psi}_{WL}$ is

$$\begin{aligned} W_\rho(1-\alpha) &= P_\rho\left(\widehat{\psi}_{WL} \le \psi\right) \\ &= \int_0^1 b(w) F_g\left(F^{-1}_{f(\widehat{\rho}(w))}(1-\alpha)\sqrt{ga_1(\rho)w + ga_2(\rho)(1-w)}\right) dw \ . \end{aligned}$$

When $\rho = 0$ or $\rho = 1$ and for any (m, n) one finds that the coverage probabilities are exactly equal to the nominal values $1-\alpha$, i.e., $K_\rho(1-\alpha) = W_\rho(1-\alpha) = 1-\alpha$. This is seen most directly from the fact that in these cases $T \sim F_{m-1}$ and $T \sim F_{n-1}$, respectively.

Figure 3 displays the exact coverage probabilities $G_\rho(.95)$ and $W_\rho(.95)$ for equal sample sizes $m = n = 2, 3, 5$ as a function of $\rho \in [0, .5]$. The full graph is symmetric around $\rho = .5$ for $m = n$. It is seen that both procedures are highly accurate even for small samples. Mostly the double bootstrap based bounds are slightly more accurate than Welch's method. However, for ρ near zero or one there is a reversal. Note how fast the curve reversal smoothes out as the sample sizes increase. Figure 4 shows the rate at which the maximum coverage error for both procedures tends to zero for $m = n = 2, \ldots, 10, 15, 20, 30, 40, 50$. It confirms the rate results given by Beran (1988). The approximate asymptotes are the lines going through $(0, 0)$ and the last point, corresponding to $m = n = 50$. It seems plausible that the true asymptotes actually coincide.

It may be of interest to find out what effect a further bootstrap iteration would have on the exact coverage rate. The formulas for these coverage rates are analogous to the previous ones with $G^{-1}_{\widehat{\rho}(w)}(1-\alpha)$ and $F^{-1}_{f(\widehat{\rho}(w))}(1-\alpha)$ replaced by appropriate iterated inverses, adding considerably to the complexity of numerical calculations. We conjecture that such an iteration will increase the number of oscillations in the coverage curve. This may then explain why further iterations may lead to highly irregular coverage behavior.

Figure 3: Coverage Probabilities of 95% Lower Bounds in the Behrens-Fisher Problem

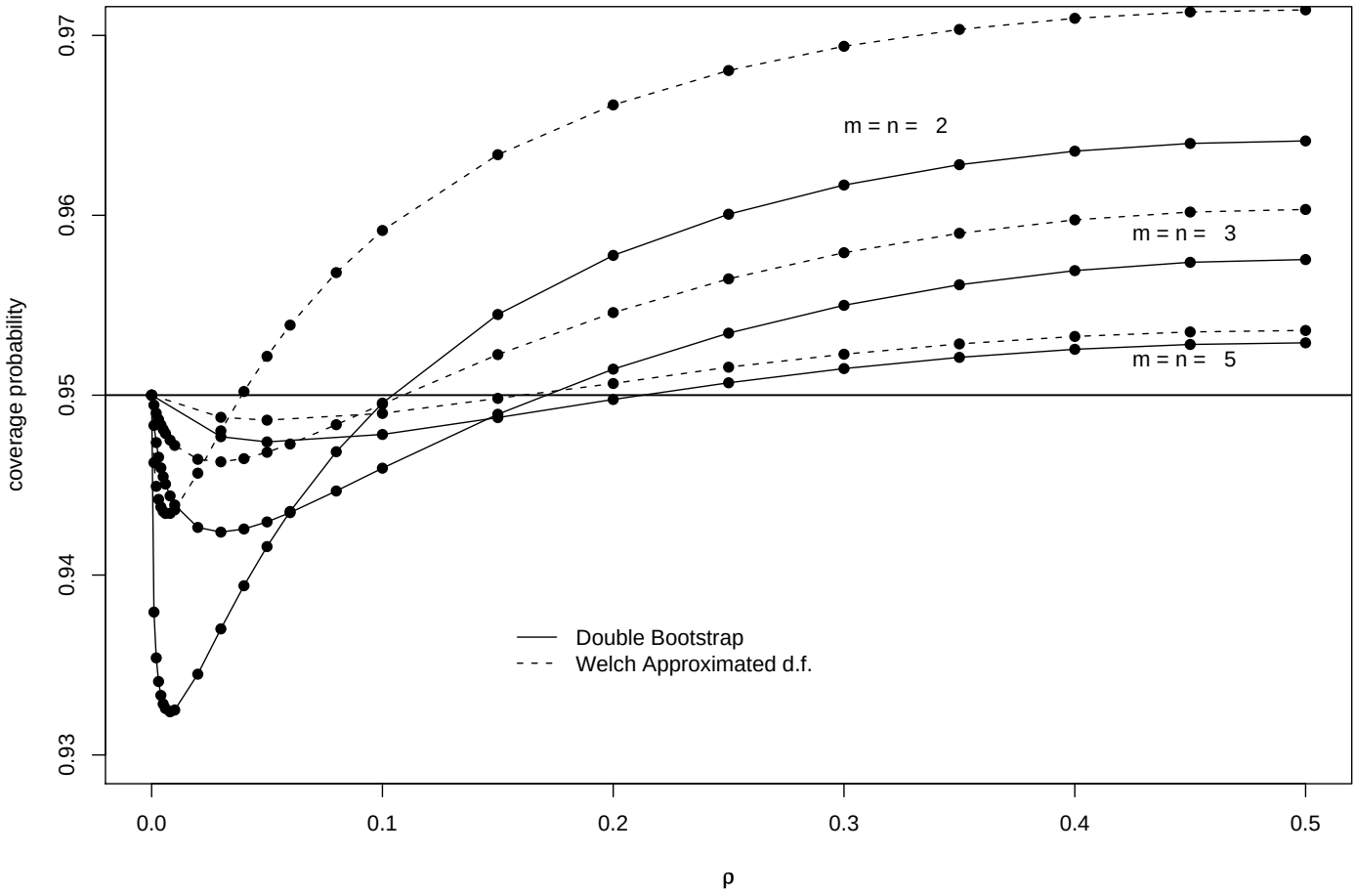
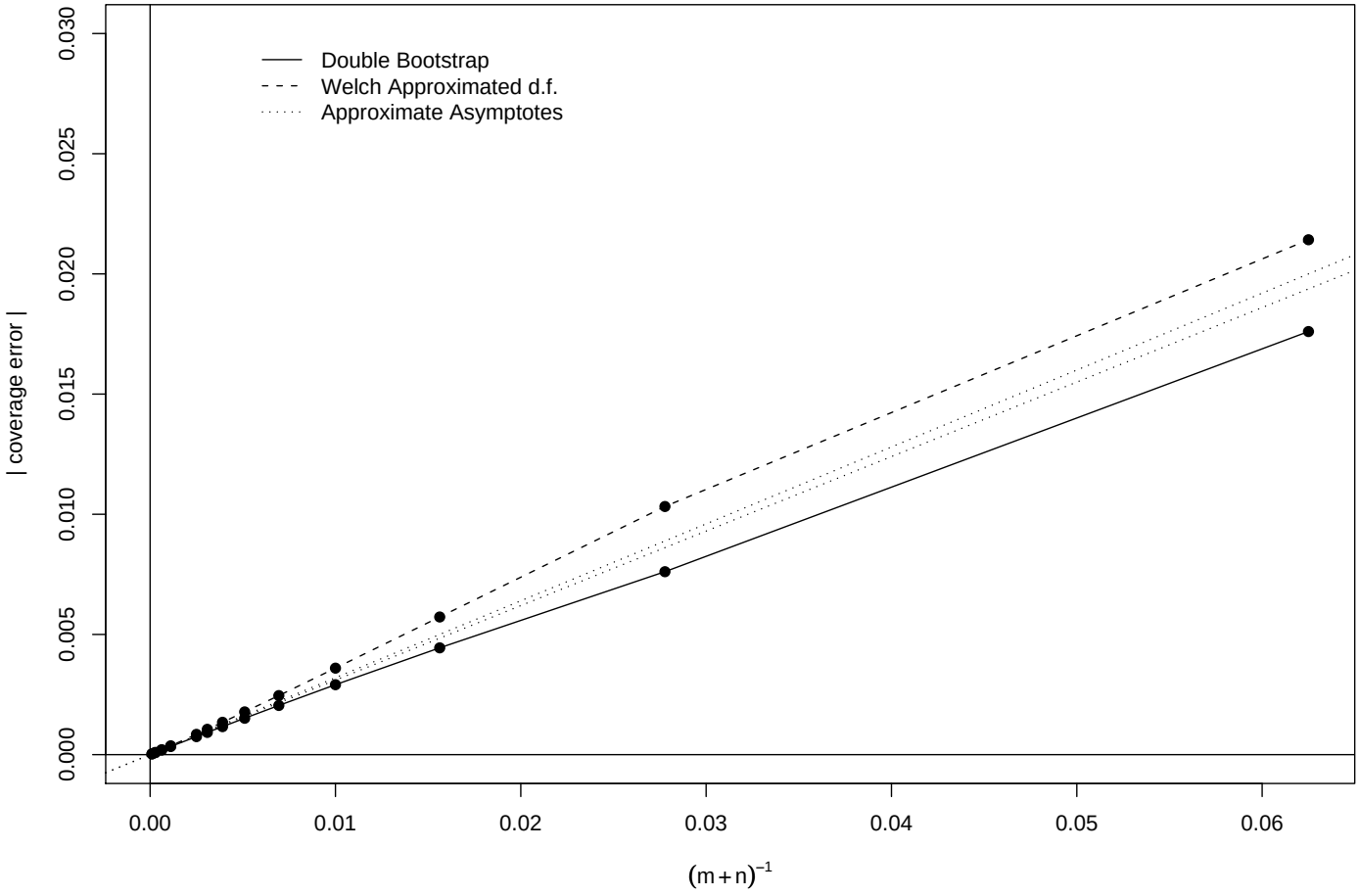


Figure 4: Maximum Coverage Error of 95% Lower Bounds in the Behrens-Fisher Problem



5.3 A Case Study

In this section we examine the small sample performance of various bootstrap methods in the context of Example 2. In particular, we examine the situation of obtaining confidence bounds for a normal percentile $\psi = \mu + z_p\sigma$. Using the notation introduced in Section 5.2.4 we take $\widehat{\theta} = (\widehat{\psi}, \widehat{\sigma})$ as estimate of $\theta = (\psi, \eta) = (\psi, \sigma)$, with

$$\widehat{\psi} = \bar{X} + ks \quad \text{and} \quad \widehat{\sigma} = rs$$

for some known constants k and $r > 0$. We will make repeated use of the following expression for the bootstrap distribution of $\widehat{\psi}^\star$

$$P_{\widehat{\psi},\widehat{\sigma}}\left(\widehat{\psi}^\star \le x\right) = D_{\widehat{\psi},\widehat{\sigma}}(x) = G_{n-1,\sqrt{n}(\widehat{\psi}-x)/\widehat{\sigma}-z_p\sqrt{n}}(-k\sqrt{n}) \;, \tag{6}$$

which follows from Equation (5) after replacing parameters by estimates. In the following subsections we will examine the true small sample coverage probabilitites of confidence bounds and equal tailed intervals for Efron's and Hall's percentile methods, for the bias corrected percentile method, the percentile-t method, and the various double bootstrap methods.

5.3.1 Efron's Percentile Method

According to Efron's percentile method we take the α-quantile of the bootstrap distribution of $\widehat{\psi}^\star$ as $100(1-\alpha)\%$ lower bound $\widehat{\psi}_L$ for ψ. The $(1-\alpha)$-quantile serves as the corresponding upper bound $\widehat{\psi}_U$ for ψ, and together these two bounds serve as a $100(1-2\alpha)\%$ confidence interval for ψ.
If we denote by $\delta_\gamma = \delta_\gamma(k,n)$ the δ which solves

$$G_{n-1,\delta}(-k\sqrt{n}) = \gamma,$$

and using (6) we can represent the above bounds as follows

$$\widehat{\psi}_L = \widehat{\psi} - \frac{\widehat{\sigma}}{\sqrt{n}}\left(\delta_\alpha(k,n) + z_p\sqrt{n}\right) = \bar{X} + k's$$

with $k' = k - rz_p - r\delta_\alpha(k,n)/\sqrt{n}$. A corresponding expression is obtained for the upper bound

$$\widehat{\psi}_U = \widehat{\psi} - \frac{\widehat{\sigma}}{\sqrt{n}}\left(\delta_{1-\alpha}(k,n) + z_p\sqrt{n}\right) = \bar{X} + k''s$$

with $k'' = k - rz_p - r\delta_{1-\alpha}(k,n)/\sqrt{n}$.
From (5) the actual coverage probabilities of these bounds are obtained as

$$P_{\psi,\sigma}\left(\widehat{\psi}_{EL} \leq \psi\right) = G_{n-1,-z_p\sqrt{n}}\left(r\delta_\alpha(k,n) + rz_p\sqrt{n} - k\sqrt{n}\right) ,$$

$$P_{\psi,\sigma}\left(\widehat{\psi}_{EU} \geq \psi\right) = 1 - G_{n-1,-z_p\sqrt{n}}\left(r\delta_{1-\alpha}(k,n) + rz_p\sqrt{n} - k\sqrt{n}\right)$$

and

$$\begin{aligned} P_{\psi,\sigma}\left(\widehat{\psi}_{EL} \leq \psi \leq \widehat{\psi}_{EU}\right) &= G_{n-1,-z_p\sqrt{n}}\left(r\delta_\alpha(k,n) + rz_p\sqrt{n} - k\sqrt{n}\right) \\ &\quad - G_{n-1,-z_p\sqrt{n}}\left(r\delta_{1-\alpha}(k,n) + rz_p\sqrt{n} - k\sqrt{n}\right) . \end{aligned}$$

Figure 5 shows the behavior of the coverage error of the 95% lower bound (with $r = 1$ and $k = z_{.10}$) for $\psi = \mu + z_{.10}\sigma$ against the theoretical rate $1/\sqrt{n}$. The actual size of the error is quite large even for large n. Also, the $1/\sqrt{n}$ asymptote is approximated well only for moderately large n, say $n \geq 20$. Figure 6 shows the corresponding result for the upper bound. Note that the size of the error is substantially smaller here. Finally, Figure 7 shows the coverage error of the 95% equal tailed confidence interval for ψ against the theoretical rate of $1/n$. The asymptote is reasonably approximated for much smaller n here.

5.3.2 Hall's Percentile Method

According to Hall's percentile method we take

$$\widehat{\psi}_{HL} = \widehat{\psi} - x^\star_{1-\alpha}$$

as $100(1-\alpha)\%$ level lower bound for ψ. Here $x^\star_{1-\alpha}$ is the $(1-\alpha)$-quantile of the bootstrap distribution for $\widehat{\psi}^\star - \widehat{\psi}$. The corresponding $100(1-\alpha)\%$ level upper bound is

$$\widehat{\psi}_{HU} = \widehat{\psi} - x^\star_\alpha ,$$

and jointly these two bounds serve as a $100(1-2\alpha)\%$ confidence interval for ψ.
From Equation (6) we obtain

$$P_{\widehat{\psi},\widehat{\sigma}}\left(\widehat{\psi}^\star - \widehat{\psi} \leq x\right) = G_{n-1,-x\sqrt{n}/\widehat{\sigma}-z_p\sqrt{n}}(-k\sqrt{n}) .$$

Thus we have

$$x^\star_{1-\alpha} = -\widehat{\sigma}\left(\delta_{1-\alpha}(k,n)/\sqrt{n} + z_p\right)$$

Figure 5: Coverage Error of Lower Confidence Bounds Using Efron's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$

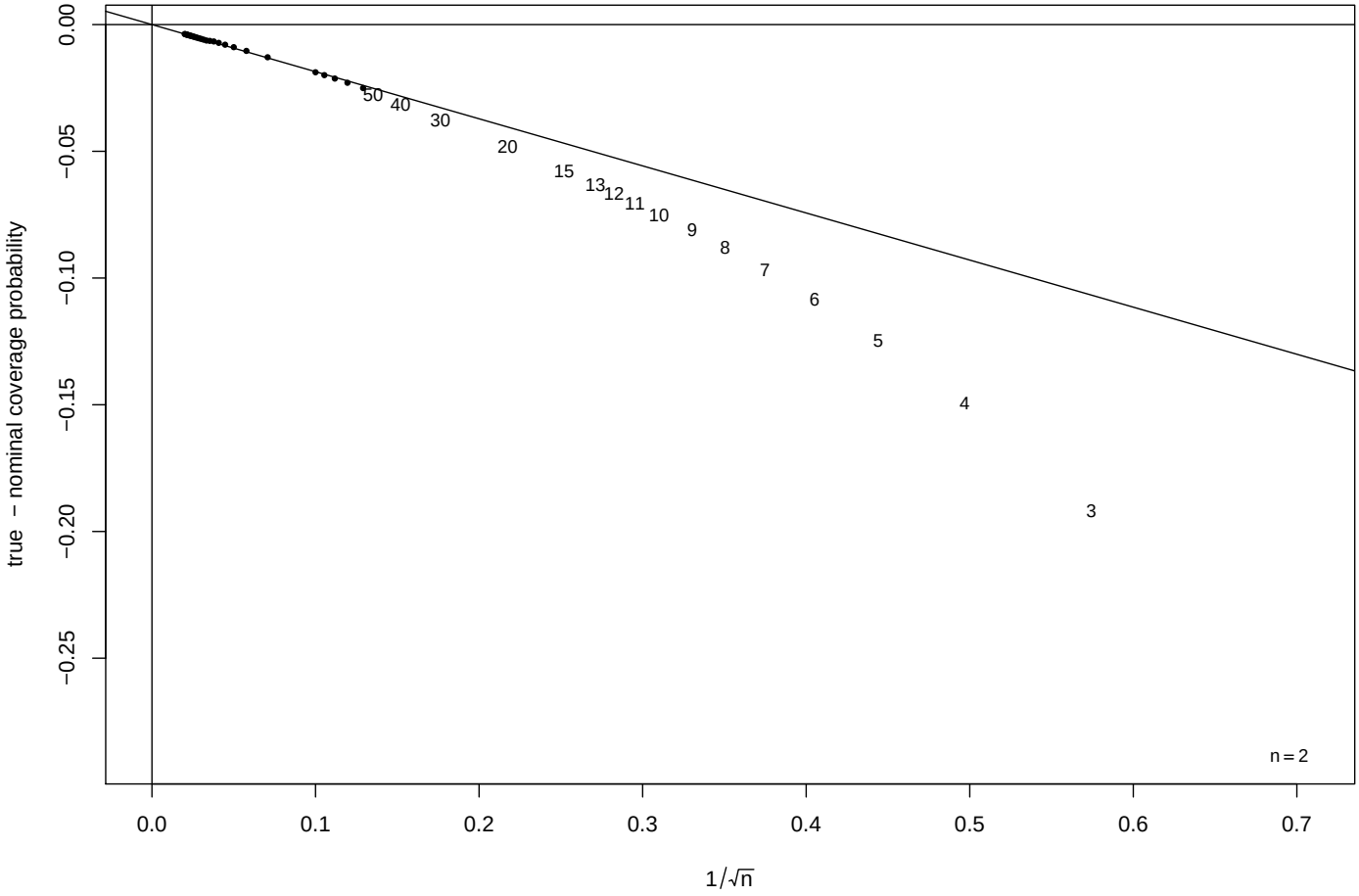


Figure 6: Coverage Error of Upper Confidence Bounds Using Efron's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$

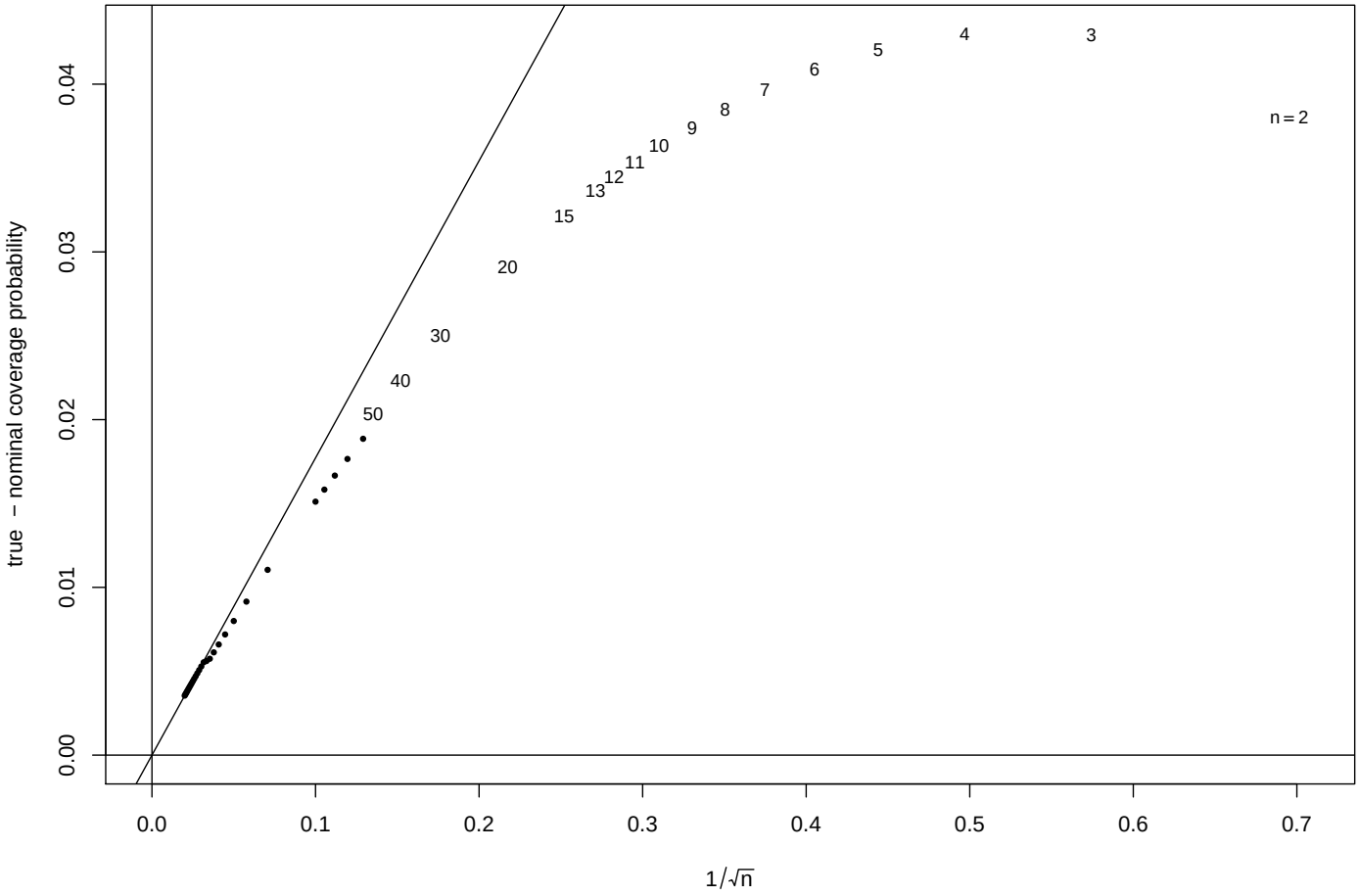
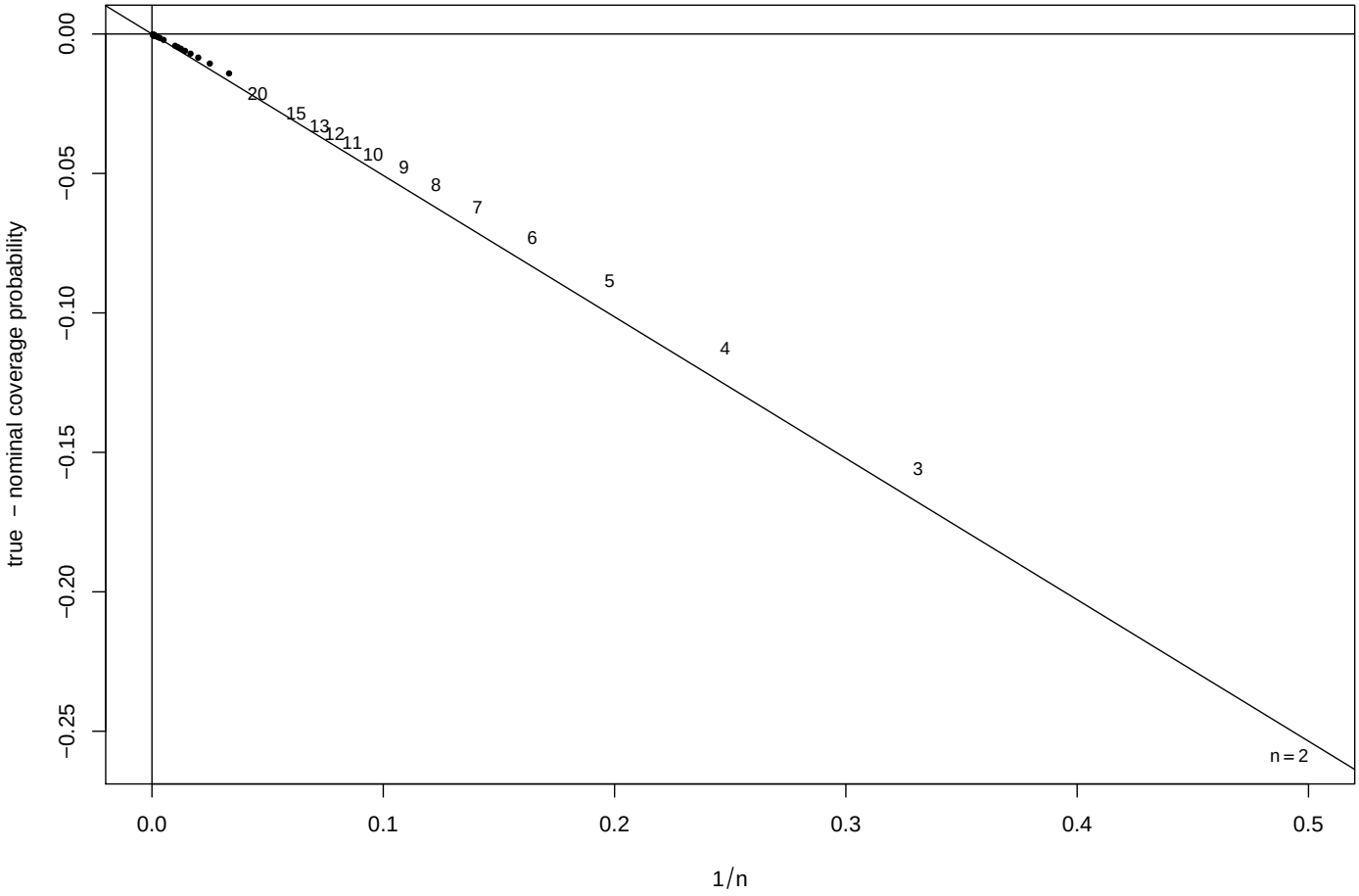


Figure 7: Coverage Error of Confidence Intervals Using Efron's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$



and thus

$$\begin{aligned}\widehat{\psi}_{HL} &= \widehat{\psi} + \widehat{\sigma}\left(\delta_{1-\alpha}(k,n)/\sqrt{n} + z_p\right) \\ &= \bar{X} + s\left(k + rz_p + r\delta_{1-\alpha}(k,n)/\sqrt{n}\right) = \bar{X} + k's\end{aligned}$$

with $k' = k + rz_p + r\delta_{1-\alpha}(k,n)/\sqrt{n}$.
From Equation 5 the actual coverage probability of $\widehat{\psi}_{HL}$ is

$$P_{\psi,\sigma}\left(\widehat{\psi}_{HL} \leq \psi\right) = G_{n-1,-z_p\sqrt{n}}(-k\sqrt{n} - rz_p\sqrt{n} - r\delta_{1-\alpha}(k,n)) \ .$$

Similarly one finds as actual coverage for the upper bound

$$P_{\psi,\sigma}\left(\widehat{\psi}_{HU} \geq \psi\right) = 1 - G_{n-1,-z_p\sqrt{n}}(-k\sqrt{n} - rz_p\sqrt{n} - r\delta_{\alpha}(k,n))$$

and for the equal tailed interval

$$\begin{aligned}P_{\psi,\sigma}\left(\widehat{\psi}_{HL} \leq \psi \leq \widehat{\psi}_{HU}\right) &= G_{n-1,-z_p\sqrt{n}}(-k\sqrt{n} - rz_p\sqrt{n} - r\delta_{1-\alpha}(k,n)) \\ &\quad - G_{n-1,-z_p\sqrt{n}}(-k\sqrt{n} - rz_p\sqrt{n} - r\delta_{\alpha}(k,n)) \ .\end{aligned}$$

Figures 8-10 show the qualitative behavior of the coverage error of these these bounds when using $k = z_{.10}$, $r = 1$, and $\gamma = .95$. The error is moderately improved over that of Efron's percentile method but again sample sizes need to be quite large before the theoretical asymptotic behavior takes hold. A clearer comparison between Hall's and Efron's percentile methods can be seen in Figures 11 and 12

5.3.3 Bias Corrected Percentile Method

The respective $100(1-\alpha)\%$ lower and upper confidence bounds by the bias corrected bootstrap method are defined as

$$\widehat{\psi}_{bcL} = D^{-1}_{\widehat{\psi},\widehat{\sigma}}\left(\Phi\left(2u_0 + z_\alpha\right)\right)$$

and

$$\widehat{\psi}_{bcU} = D^{-1}_{\widehat{\psi},\widehat{\sigma}}\left(\Phi\left(2u_0 + z_{1-\alpha}\right)\right) \ ,$$

where

$$u_0 = \Phi^{-1}\left(D_{\widehat{\psi},\widehat{\sigma}}(\widehat{\psi})\right) = \Phi^{-1}\left(G_{n-1,-z_p\sqrt{n}}(-k\sqrt{n})\right)$$

Figure 8: Coverage Error of Lower Confidence Bounds Using Hall's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$

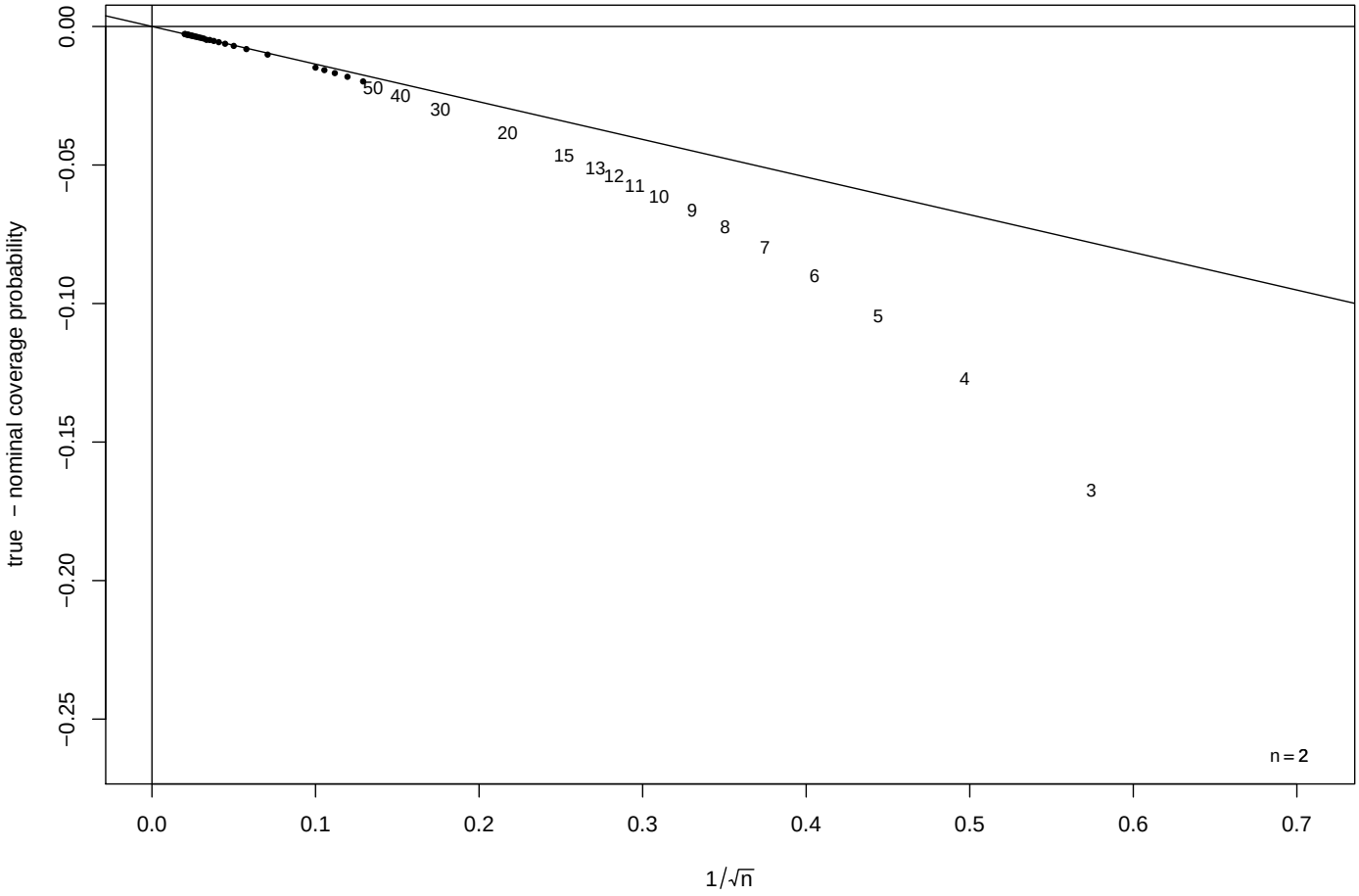


Figure 9: Coverage Error of Upper Confidence Bounds Using Hall's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$

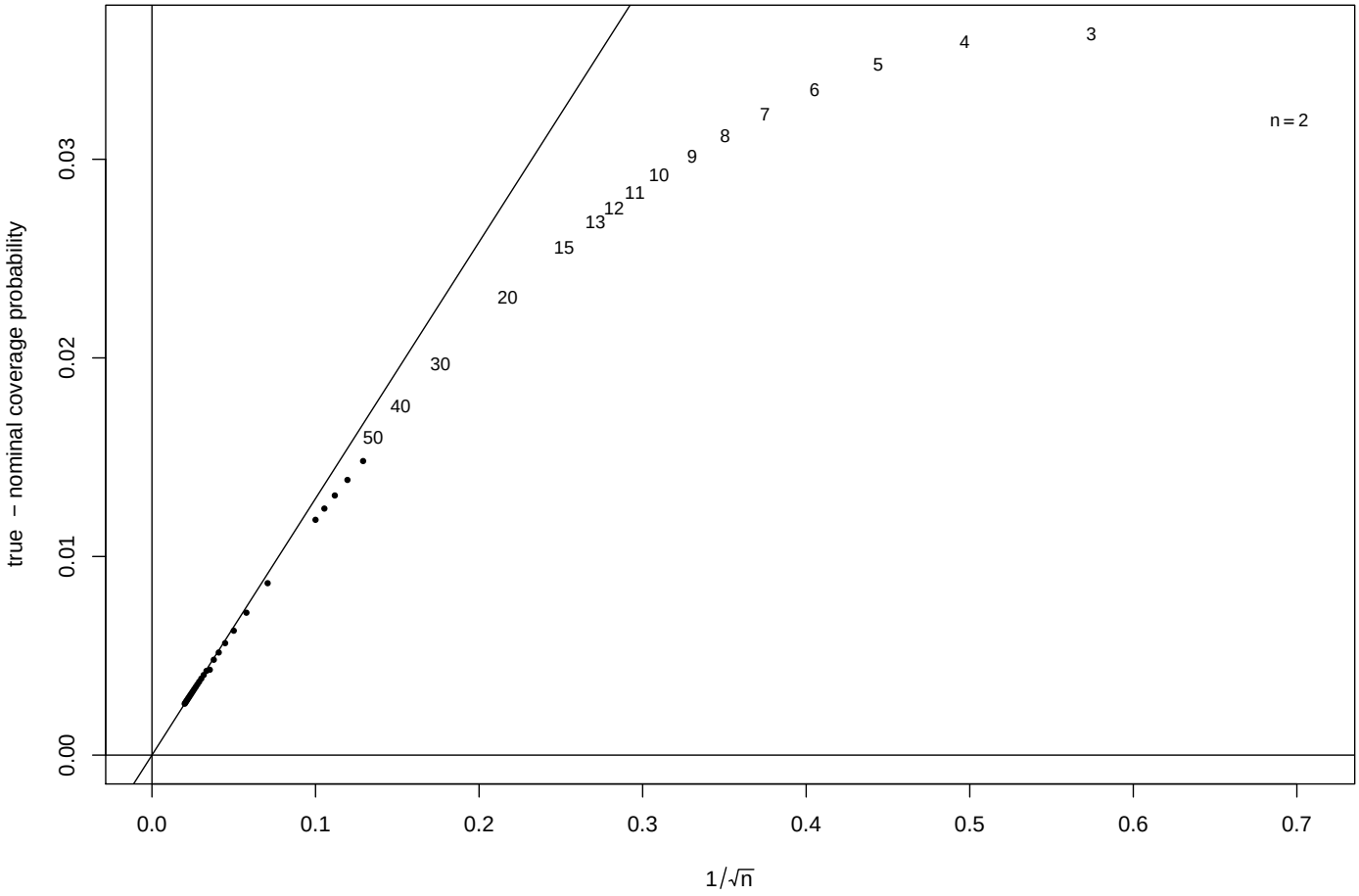
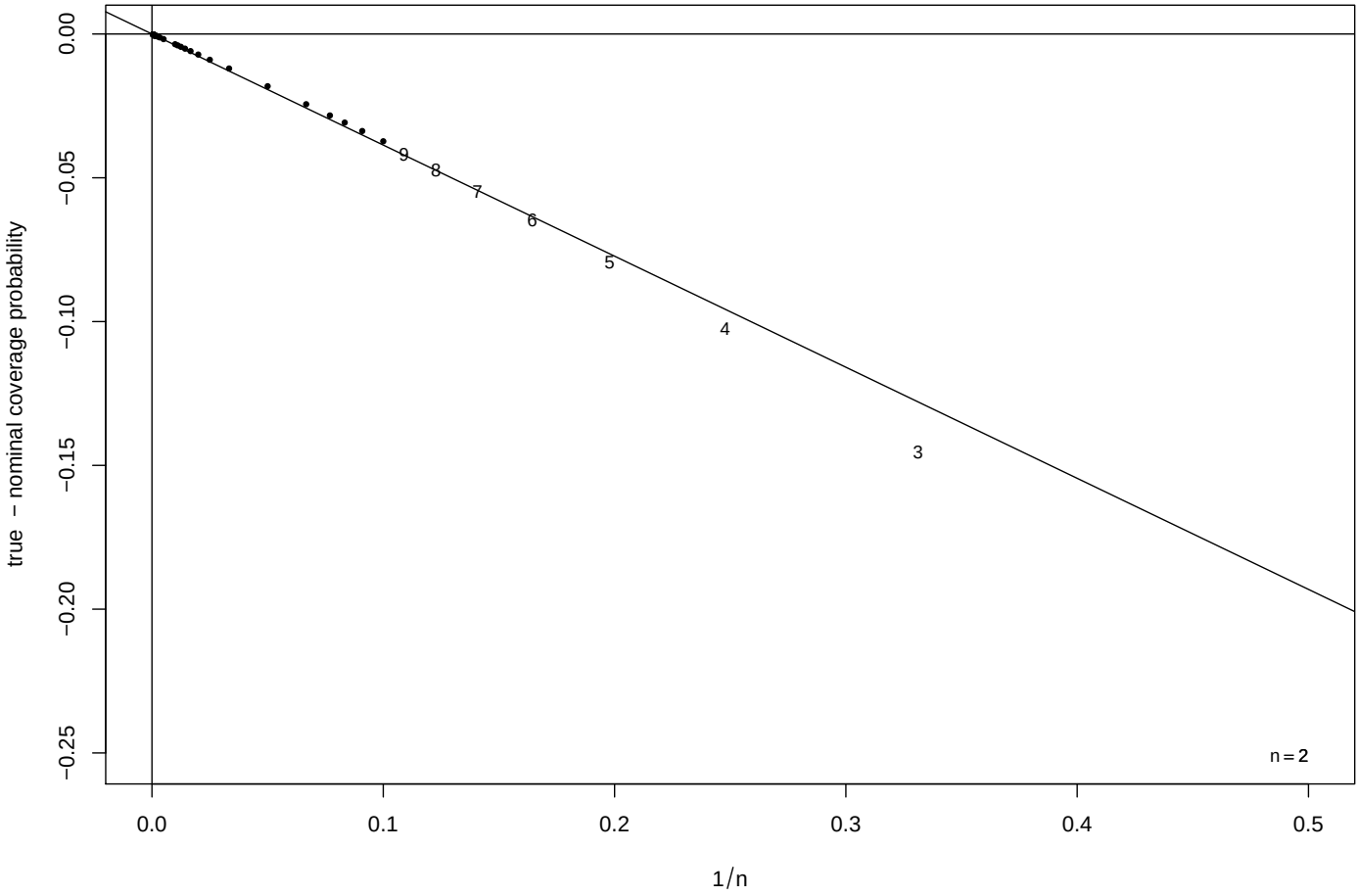


Figure 10: Coverage Error of Confidence Intervals Using Hall's Percentile Method with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$



and Φ denotes the standard normal distribution function. When $u_0 = 0$ these bounds reduce to Efron's percentile bounds. Since $x = \widehat{\psi}_{bcU}$ solves

$$D_{\widehat{\psi},\widehat{\sigma}}(x) = G_{n-1,\sqrt{n}(\widehat{\psi}-x)/\widehat{\sigma}-z_p\sqrt{n}}(-k\sqrt{n}) = \Phi\left(2u_0 + z_{1-\alpha}\right) = \gamma(1-\alpha)$$

we can express $\widehat{\psi}_{bcU}$ as follows

$$\widehat{\psi}_{bcU} = \widehat{\psi} - \widehat{\sigma}\left(\delta_{\gamma(1-\alpha)}(k,n) + z_p\sqrt{n}\right)/\sqrt{n} = \bar{X} + sk_U$$

with $k_U = k - rz_p - r\delta_{\gamma(1-\alpha)}(k,n)/\sqrt{n}$.
The corresponding expression for the bias corrected lower bound is

$$\widehat{\psi}_{bcL} = \widehat{\psi} - \widehat{\sigma}\left(\delta_{\gamma(\alpha)}(k,n) + z_p\sqrt{n}\right)/\sqrt{n} = \bar{X} + sk_L$$

with

$$\gamma(\alpha) = \Phi\left(2u_0 + z_\alpha\right) \quad \text{and} \quad k_L = k - rz_p - r\delta_{\gamma(\alpha)}(k,n)/\sqrt{n}\ .$$

From (5) the respective exact coverage probabilities are obtained as

$$P_{\psi,\sigma}\left(\widehat{\psi}_{bcL} \le \psi\right) = G_{n-1,-z_p\sqrt{n}}(-k_L\sqrt{n})\ ,$$

$$P_{\psi,\sigma}\left(\widehat{\psi}_{bcU} \ge \psi\right) = 1 - G_{n-1,-z_p\sqrt{n}}(-k_U\sqrt{n})\ ,$$

and for the equal tailed interval

$$\begin{aligned} P_{\psi,\sigma}\left(\widehat{\psi}_{bcL} \le \psi \le \widehat{\psi}_{bcU}\right) &= G_{n-1,-z_p\sqrt{n}}(-k_L\sqrt{n}) \\ &= -G_{n-1,-z_p\sqrt{n}}(-k_U\sqrt{n})\ . \end{aligned}$$

Figure 11 compares the qualitative behavior of upper and lower confidence bounds by the bias corrected percentile method, Efron's percentile method, and Hall's percentile method. The comparison is again made against the theoretical rate $1/\sqrt{n}$, which is appropriate for all three methods. Bias correction appears to improve over both the other percentile methods. However, for small sample sizes the magnitude of the actual coverage error is the more dominant feature. The asymptotes are again approached only for large n. The differences in coverage error become small relative to the actual coverage error for large n. Figure 12 portrays the corresponding coverage properties for equal tailed confidence intervals against the relevant rate of $1/n$. The above observations apply here as well.

Figure 11: Coverage Error of Confidence Bounds Comparing Percentile Methods and Bias Correction with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in Normal Population, $p = .1$ and Confidence $\gamma = .95$

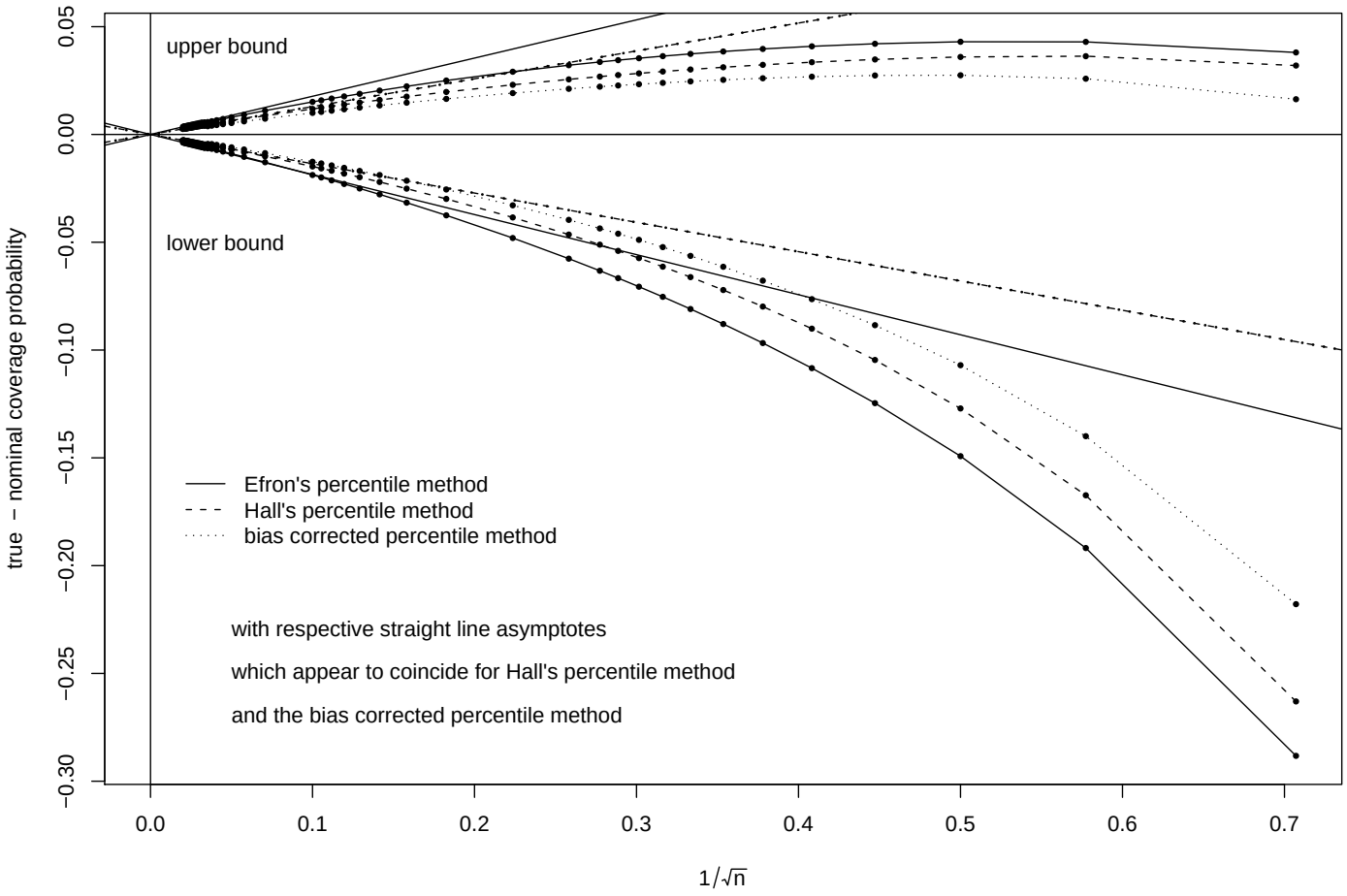
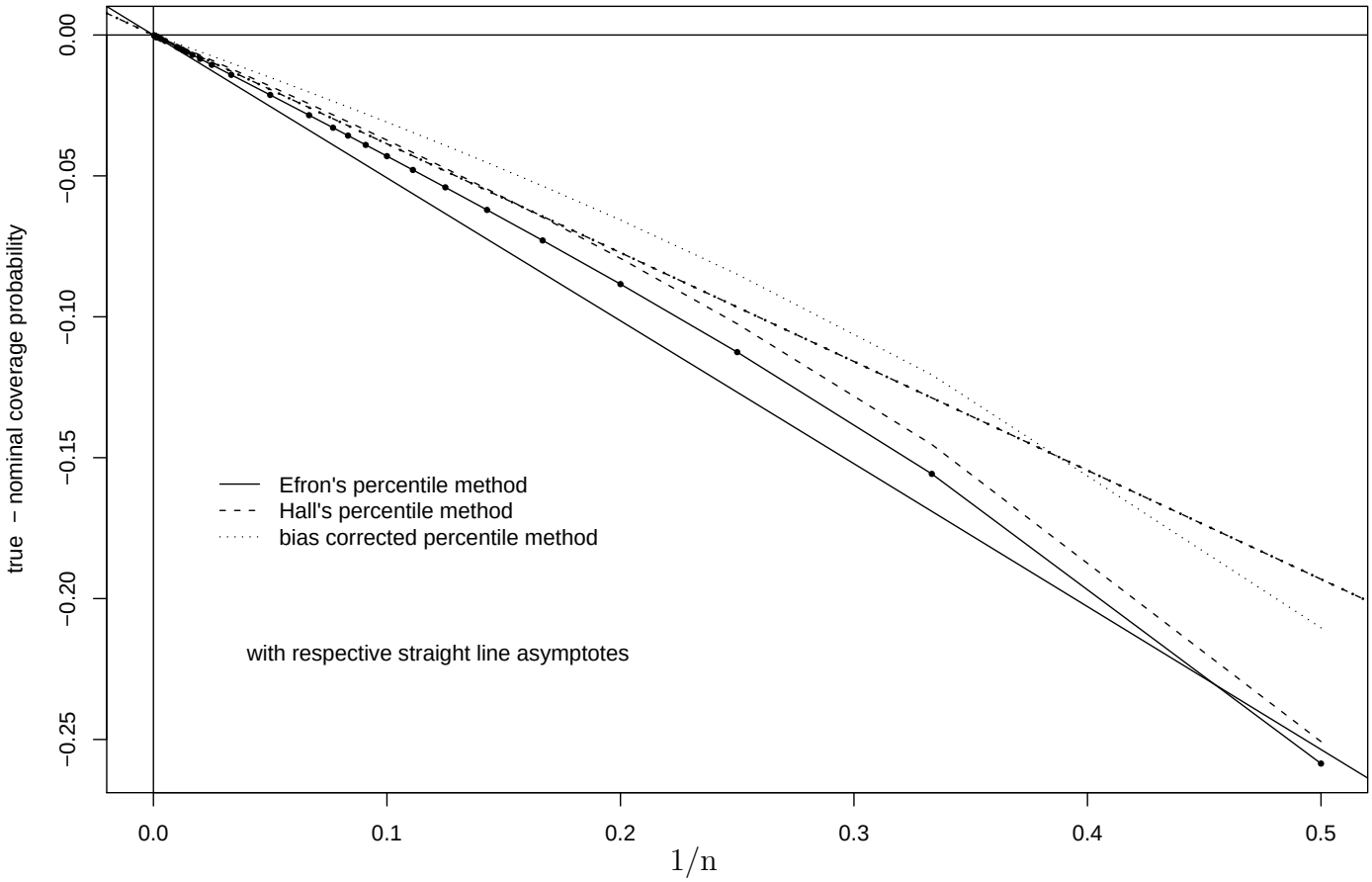


Figure 12: Coverage Error of Confidence Intervals Comparing Percentile Methods and Bias Correction with $\bar{X} + z_p\, s$ Estimating $\psi = \mu + z_p\, \sigma$ in a Normal Population, $p = .1$ and Confidence $\gamma = .95$



5.3.4 Percentile-t and Double Bootstrap Methods

Using $T = (\widehat{\psi} - \psi)/\widehat{\sigma}$ as the Studentized ratio in the percentile-t method will result in the classical confidence bounds and thus there will be no coverage error. This arises because T is an exact pivot.
If we take $R = \widehat{\psi} - \psi$ as a root in Beran's prepivoting method, we again arrive at the same classical confidence bounds. This was already pointed out in Section 5.2.3 and is due to the fact that the distribution of R only depends on the nuisance parameter σ.
The automatic double bootstrap also arrives at the classical confidence bounds as was already examined in Section 5.2.4. Thus the automatic double bootstrap succeeds here without having to make a choice of scale estimate for Studentization or of a root for prepivoting.

5.4 References

Aspin, A.A. (1949). "Tables for use in comparisons whose accuracy involves two variances, separately estimated (with an Appendix by B.L. Welch)." *Biometrika* **36**, 290-296.

Bain, L.J. (1987). *Statistical Analysis of Reliability and Life-Testing Models, Theory and Methods.* Marcel Dekker, Inc., New York.

Beran, R. (1987). "Prepivoting to reduce level error of confidence sets." *Biometrika* **74**, 457-468.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements." *J. Amer. Statist. Assoc.* **83**, 687-697.

Diaconis, P. and Efron, B. (1983a). "Computer intensive methods in statistics." *Sci. Amer.* **248**, 116-130.

Diaconis, P. and Efron, B. (1983b). "Statistik per Computer: der Münchhausen-Trick." *Spektrum der Wissenschaft*, Juli 1983, 56-71. German translation of Diaconis, P. and Efron, B. (1983a), introducing the German term Münchhausen for bootstrap.

DiCiccio, T.J. and Romano, J.P. (1988). "A review of bootstrap confidence intervals." (With discussion) *J. Roy. Statist. Soc. Ser. B* **50**, 338-354.

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife." *Ann. Statist.* **7**, 1-26.

Efron, B. (1981). "Nonparametric standard errors and confidence intervals." (With discussion) *Canad. J. Statist.* **9**, 139-172.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. (1987). "Better bootstrap confidence intervals." (With discussion) *J. Amer. Statist. Assoc.* **82**, 171-200.

Fleiss, J.L. (1971). "On the distribution of a linear combination of independent chi squares." *J. Amer. Statist. Assoc.* **66**, 142-144.

Hall, P. (1988a). "Theoretical comparison of bootstrap confidence intervals." (With discussion) *Ann. Statist.* **16**, 927-985.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses, Second Edition*, John Wiley & Sons, New York.

Loh, W. (1987). "Calibrating confidence coefficients." *J. Amer. Statist. Assoc.* **82**, 155-162.

Reid, N. (1981). Discussion of Efron (1981).

Scholz, F.-W. (1994). "On exactness of the parametric double bootstrap." it Statistica Sinica, **4**, 477-492.

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