Chapter 18
Bootstrap and Subsampling Methods

18.1 Introduction

The bootstrap, subsampling, and other resampling methods provide methods for inference, especially in problems where large-sample approximations are not tractable. Such methods are not foolproof and require mathematical justification. In this chapter, fundamental properties of these methods are developed.

In Section 18.2 we first review some basic constructions of confidence regions and tests, which derive from the limiting distribution of an estimator or test sequence. This serves to motivate the bootstrap construction studied in Section 18.3; the bootstrap method offers a powerful approach to approximating the sampling distribution of a given statistic or estimator. The emphasis here is to find methods that control the level constraint, at least asymptotically. Like the randomization construction, the bootstrap approach will be asymptotically efficient if the given statistic is chosen appropriately; for example, see Theorem 18.3.2 and Corollary 18.3.1.

While the bootstrap is quite general, how does it compare in situations when other large-sample approaches apply as well? In Section 18.4, we provide some support to the claim that the bootstrap approach can improve upon methods that rely on a normal approximation. The use of the bootstrap in the context of hypothesis testing is studied in Section 18.5.

While the bootstrap method is quite broadly applicable, in some situations, it can be inconsistent. A more general approach based on subsampling is presented in Section 18.7. Together, these approaches serve as valuable tools for inference without having to make strong assumptions about the underlying distribution.

18.2 Basic Large-Sample Approximations

In the previous section, it was shown how permutation and randomization tests can be used in certain problems where the randomization hypothesis holds. Unfortunately, randomization tests only apply to a restricted class of problems. In this section,

E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics, https://doi.org/10.1007/978-3-030-70578-7_18

we discuss some generally used asymptotic approaches for constructing confidence regions or hypothesis tests based on data $X = X^n$. In what follows, $X^n = (X_1, \dots, X_n)$ is typically a sample of n i.i.d. random variables taking values in a sample space S and having unknown probability distribution P, where P is assumed to belong to a certain collection $\mathbf{P}$ of distributions. Even outside the i.i.d. case, we think of the data X^n as coming from a model indexed by the unknown probability mechanism P. The collection $\mathbf{P}$ may be a parametric model indexed by a Euclidean parameter, but we will also consider nonparametric models.

We shall be interested in inferences concerning some parameter $\theta(P)$. By the usual duality between the construction of confidence regions and hypothesis tests, we can restrict the discussion to the construction of confidence regions. Let the range of θ be denoted by Θ, so that

$$\Theta = \{\theta(P) : P \in \mathbf{P}\} .$$

Typically, Θ is a subset of the real line, but we also consider more general parameters. For example, the problem of estimating the entire cumulative distribution function (c.d.f.) of real-valued observations may be treated, so that Θ is an appropriate function space.

This leads to considering a *root* $R_n(X^n, \theta(P))$, a term first coined by Beran (1984), which is just some real-valued functional depending on both X^n and $\theta(P)$. The idea is that a confidence interval for $\theta(P)$ could be constructed if the distribution of the root were known. For example, an estimator $\hat{\theta}_n$ of a real-valued parameter $\theta(P)$ might be given so that a natural choice is $R_n(X^n, \theta(P)) = [\hat{\theta}_n - \theta(P)]$, or alternatively $R_n(X^n, \theta(P)) = [\hat{\theta}_n - \theta(P)]/s_n$, where s_n is some estimate of the standard deviation of $\hat{\theta}_n$.

When $\mathbf{P}$ is suitably large so that the problem is nonparametric in nature, a natural construction for an estimator $\hat{\theta}_n$ of $\theta(P)$ is the plug-in estimator $\hat{\theta}_n = \theta(\hat{P}_n)$, where $\hat{P}_n$ is the empirical distribution of the data, defined by

$$\hat{P}_n(E) = n^{-1} \sum_{i=1}^{n} I\{X_i \in E\} .$$

Of course, this construction implicitly assumes that $\theta(\cdot)$ is defined for empirical distributions so that $\theta(\hat{P}_n)$ is at least well defined. Alternatively, in parametric problems for which $\mathbf{P}$ is indexed by a parameter ψ belonging to a subset Ψ of $\mathbb{R}^p$ so that $\mathbf{P} = \{P_\psi : \psi \in \Psi\}$, then $\theta(P)$ can be described as a functional $t(\psi)$. Hence, $\hat{\theta}_n$ is often taken to be $t(\hat{\psi}_n)$, where $\hat{\psi}_n$ is some desirable estimator of ψ, such as an efficient likelihood estimator.

Let $J_n(P)$ be the distribution of $R_n(X^n, \theta(P))$ under P, and let $J_n(\cdot, P)$ be the corresponding cumulative distribution function defined by

$$J_n(x, P) = P\{R_n(X^n, \theta(P)) \le x\}.$$

In order to construct a confidence region for $\theta(P)$ based on the root $R_n(X^n, \theta(P))$, the sampling distribution $J_n(P)$ or its appropriate quantiles must be known or estimated. Some standard methods, based on pivots and asymptotic approximations, are now briefly reviewed. Note that in many of the examples when the observations are real-valued, it is more convenient and customary to index the unknown family of distributions by the cumulative distribution function F rather than P. We will freely use both, depending on the situation.

18.2.1 Pivotal Method

In certain exceptional cases, the distribution $J_n(P)$ of $R_n(X^n, \theta(P))$ under P does not depend on P. In this case, the root $R_n(X^n, \theta(P))$ is called a *pivotal quantity* or a *pivot* for short. Such quantities were previously considered in Section 6.12. From a pivot, a level $1 - \alpha$ confidence region for $\theta(P)$ can be constructed by choosing constants c_1 and c_2 so that

$$P\{c_1 \le R_n(X^n, \theta(P)) \le c_2\} \ge 1 - \alpha \ . \tag{18.1}$$

Then, the confidence region

$$C_n = \{\theta \in \Theta : \ c_1 \le R_n(X^n, \theta) \le c_2\}$$

contains $\theta(P)$ with probability under P at least $1 - \alpha$. Of course, the coverage probability is exactly $1 - \alpha$ if one has equality in (18.1).

Classical examples where confidence regions may be formed from a pivot are the following.

Example 18.2.1 (**Location and Scale Families**) Suppose we are given an i.i.d. sample $X^n = (X_1, \ldots, X_n)$ of n real-valued random variables, each having a distribution function of the form $F[(x - \theta)/\sigma]$, where F is known, θ is a location parameter, and σ is a scale parameter. More generally, suppose $\hat{\theta}_n$ is location and scale equivariant in the sense that

$$\hat{\theta}_n(aX_1 + b, \ldots, aX_n + b) = a\hat{\theta}_n(X_1, \ldots, X_n) + b \ ;$$

also suppose $\hat{\sigma}_n$ is location invariant and scale equivariant in the sense that

$$\hat{\sigma}_n(aX_1 + b, \ldots, aX_n + b) = |a|\hat{\sigma}_n(X_1, \ldots, X_n) \ .$$

Then, the root $R_n(X^n, \theta(P)) = n^{1/2}[\hat{\theta}_n - \theta(P)]/\hat{\sigma}_n$ is a pivot (Problem 18.1). For example, in the case where F is the standard normal distribution function, $\hat{\theta}_n$ is the sample mean and $\hat{\sigma}_n^2$ is the usual unbiased estimate of variance, R_n has a t-distribution with $n - 1$ degrees of freedom. For another example, if $\hat{\sigma}_n$ is location invariant and scale equivariant, then $\hat{\sigma}_n/\sigma$ is also a pivot, since its distribution will not depend on θ or σ, but will of course depend on F. When F is not normal, exact distribution theory may be difficult, but one may resort to Monte Carlo simulation of $J_n(P)$ (discussed below). This example can be generalized to a class of parametric problems where group invariance considerations apply, and pivotal quantities lead to equivariant confidence sets; see Section 6.12 and Problems 6.71–6.74. ■

Example 18.2.2 (**Kolmogorov–Smirnov Confidence Bands**) Suppose that $X^n = (X_1, \cdots, X_n)$ is a sample of n real-valued random variables having a distribution function F. For a fixed value of x, a (pointwise) confidence interval for $F(x)$ can be based on the empirical distribution function $\hat{F}_n(x)$, by using the fact that $n\hat{F}_n(x)$ has a binomial distribution with parameters n and $F(x)$. The goal now is to construct a uniform or simultaneous confidence band for $\theta(F) = F$, so that it is required to find a set of distribution functions containing the true $F(x)$ for all x (or uniformly in x) with coverage probability $1 - \alpha$. Toward this end, consider the root

$$R_n(X^n, F) = n^{1/2} \sup_x |\hat{F}_n(x) - F(x)|.$$

Recall that, if F is continuous, then the distribution of $R_n(X^n, F)$ under F does not depend on F and so $R_n(X^n, F)$ is a pivot (Section 6.13 and Problem 11.68). As discussed in Sections 6.13 and 16.2, the finite-sample quantiles of this distribution have been tabled. Without the assumption that F is continuous, the distribution of $R_n(X^n, F)$ under F does depend on F, both in finite samples and asymptotically. ■

In general, if $R_n(X^n, \theta(P))$ is a pivot, its distribution may not be explicitly computable or have a known tractable form. However, since there is only one distribution that needs to be known (and not an entire family indexed by P), the problem is much simpler than if the distribution depends on P. One can resort to Monte Carlo simulation to approximate this distribution to any desired level of accuracy, by simulating the distribution of $R_n(X^n, \theta(P))$ under P for any choice of P in $\mathbf{P}$. For further details, see Example 11.4.3.

18.2.2 Asymptotic Pivotal Method

In general, the above construction breaks down because $R_n(X^n, \theta(P))$ has a distribution $J_n(P)$ which depends on the unknown probability distribution P generating the data. However, it is then sometimes the case that $J_n(P)$ converges weakly to a limiting distribution J which is independent of P. In this case, the root (sequence) $R_n(X^n, \theta(P))$ is called an *asymptotic pivot*, and then the quantiles of J may be used to construct an asymptotic confidence region for $\theta(P)$.

Example 18.2.3 (**Parametric Models**) Suppose $X^n = (X_1, \ldots, X_n)$ is a sample from a model $\{P_\theta,\ \theta \in \Omega\}$, where Ω is a subset of $\mathbb{R}^k$. To construct a confidence region for θ, suppose $\hat{\theta}_n$ is an efficient likelihood estimator (as discussed in Section 14.4), satisfying

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta)) ,$$

where $I(\theta)$ is the Fisher Information matrix, assumed continuous. Then, the root (expressed as a function of θ rather than P_θ)

$$R_n(X^n, \theta) = n(\hat{\theta}_n - \theta)^\top I(\hat{\theta}_n)(\hat{\theta}_n - \theta)$$

is an asymptotic pivot. The limiting distribution is the χ^2_k, the Chi-squared distribution with k degrees of freedom, and the resulting confidence region is Wald's confidence ellipsoid introduced in Section 14.4.2. Alternatively, let

$$\tilde{R}_n(X^n, \theta) = \frac{\sup_{\beta \in \Omega} L_n(\beta)}{L_n(\theta)} ,$$

where $L_n(\theta)$ is the likelihood function (14.56). As discussed in Section 14.4.2, under regularity conditions, $2 \log \tilde{R}_n(X^n, \theta)$ is asymptotically χ^2_k, in which case $\tilde{R}_n(X^n, \theta)$ is an asymptotic pivot. ■

Example 18.2.4 (**Nonparametric Mean**) Suppose $X^n = (X_1, \ldots, X_n)$ is a sample of n real-valued random variables having distribution function F, and we wish to construct a confidence interval for $\theta(F) = E_F(X_i)$, the mean of the observations. Assume X_i has a finite nonzero variance $\sigma^2(F)$. Let the root R_n be the usual t-statistic defined by $R_n(X^n, \theta(F)) = n^{1/2}[\bar{X}_n - \theta(F)]/S_n$, where $\bar{X}_n$ is the sample mean and S_n^2 is the (unbiased version of the) sample variance. Then, $J_n(F)$ converges weakly to $J = N(0, 1)$, and so the t-statistic is an asymptotic pivot. ■

18.2.3 Asymptotic Approximation

The pivotal method assumes the root has a distribution $J_n(P)$ which does not depend on P, while the asymptotic pivotal method assumes the root has an asymptotic distribution $J(P)$ which does not depend on P. More generally, $J_n(P)$ converges to a limiting distribution $J(P)$ which depends on P, and we shall now consider this case. Suppose that this limiting distribution has a known form which depends on P, but only through some unknown parameters. For example, in the nonparametric mean example, the root $n^{1/2}[\bar{X}_n - \theta(F)]$ has the $N(0, \sigma^2(F))$ distribution, and so depends on F through the variance parameter $\sigma^2(F)$. An approximation of the asymptotic distribution is $J(\hat{P}_n)$, where $\hat{P}_n$ is some estimate of P. Typically, $J(P)$ is a normal distribution with mean zero and variance $\tau^2(P)$. The approximation then consists of a normal approximation based on an estimated variance $\tau^2(\hat{P}_n)$ which converges in probability to $\tau^2(P)$, and the quantiles of $J_n(P)$ may then be approximated by those of $J(\hat{P}_n)$. Of course, this approach depends very heavily on knowing the form of the asymptotic distribution as well as being able to construct consistent estimates of the unknown parameters upon which $J(P)$ depends. Moreover, the method essentially consists of a double approximation; first, the finite sampling distribution $J_n(P)$ is approximated by an asymptotic approximation $J(P)$, and then $J(P)$ is in turn approximated by $J(\hat{P}_n)$.

The most general situation occurs when the limiting distribution $J(P)$ has an unknown form, and methods to handle this case will be treated in the subsequent sections.

Example 18.2.5 (**Nonparametric Mean, continued**) In the previous example, consider instead the non-studentized root

$$R_n(X^n, \theta(F)) = n^{1/2}[\bar{X}_n - \theta(F)] .$$

In this case, $J_n(F)$ converges weakly to $J(F)$, the normal distribution with mean zero and variance $\sigma^2(F)$. The resulting approximation to $J_n(F)$ is the normal distribution with mean zero and variance S_n^2. Alternatively, one can estimate the variance by any consistent estimator, such as the sample variance $\sigma^2(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution function. In effect, studentizing an asymptotically normal root converts it to an asymptotic pivot, and both methods lead to the same solution. (However, the bootstrap approach in the next section treats the roots differently.) ■

Example 18.2.6 (**Binomial** p) As in Example 11.3.4, suppose X is binomial based on n trials and success probability p. Let $\hat{p}_n = X/n$. Like the previous example, the non-studentized root $n^{1/2}(\hat{p}_n - p)$ and the studentized root $n^{1/2}(\hat{p}_n - p)/[\hat{p}_n(1 - \hat{p}_n)]^{1/2}$ lead to the same approximate confidence interval given by (11.23). On the other hand, the Wilson interval (11.25) based on the root $n^{1/2}(\hat{p}_n - p)/[p(1 - p)]^{1/2}$ leads to a genuinely different solution which performs better in finite samples; see Brown et al. (2001). ■

Example 18.2.7 (**Trimmed mean**) Suppose $X^n = (X_1, \ldots, X_n)$ is a sample of n real-valued random variables with unknown distribution function F. Assume that F is symmetric about some unknown value $\theta(F)$. Let $\hat{\theta}_{n,\alpha}(X_1, \ldots, X_n)$ be the α-trimmed mean; specifically,

$$\hat{\theta}_{n,\alpha} = \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} X_{(i)} ,$$

where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the order statistics and $k = [\alpha n]$ is the greatest integer less than or equal to αn. Consider the root $R_n(X^n, \theta(F)) = n^{1/2}[\hat{\theta}_{n,\alpha} - \theta(F)]$. Then, under reasonable smoothness conditions on F and assuming $0 \le \alpha < 1/2$, it is known that $J_n(F)$ converges weakly to the normal distribution $J(F)$ with mean zero and variance $\sigma^2(\alpha, F)$, where

$$\sigma^2(\alpha, F) = \tag{18.2}$$
$$\frac{1}{(1-2\alpha)^2}\,[\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (t-\theta(F))^2 dF(t) + 2\alpha(F^{-1}(\alpha)-\theta(F))^2];$$

see Serfling (1980, p. 236). Then, a very simple first-order approximation to $J(F)$ is $J(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution. The resulting $J(\hat{F}_n)$ is just the normal distribution with mean zero and variance $\sigma^2(\alpha, \hat{F}_n)$. ■

The use of the normal approximation in the previous example hinged on the availability of a consistent estimate of the asymptotic variance. The simple expression (18.2) easily led to a simple estimator. However, a closed form expression for the asymptotic variance may not exist. A fairly general approach to estimating the variance of a statistic is provided by the *jackknife* estimator of variance, for which we refer the reader to Shao and Tu (1995, Chapter 2). However, the double approximation based on asymptotic normality and an estimate of the limiting variance may be poor. An alternative approach that more directly attempts to approximate the finite-sample distribution will be presented in the next section.

18.3 Bootstrap Sampling Distributions

18.3.1 Introduction and Consistency

In this section, the bootstrap, due to Efron (1979), is introduced as a general method to approximate a sampling distribution of a statistic or a root (discussed in Section 18.2) in order to construct confidence regions for a parameter of interest. The use of the bootstrap to approximate a null distribution in the construction of hypothesis tests will be considered later as well.

The asymptotic approaches in the previous section are not always applicable, as when the limiting distribution does not have a tractable form. Even when a root has a known limiting distribution, the resulting approximation may be poor in finite samples. The bootstrap procedure discussed in this section is an alternative, more general, direct approach to approximate the sampling distribution $J_n(P)$. An important aspect of the problem of estimating $J_n(P)$ is that, unlike the usual problem of estimation of parameters, $J_n(P)$ depends on n.

The bootstrap method consists of directly estimating the exact finite sampling distribution $J_n(P)$ by $J_n(\hat{P}_n)$, where $\hat{P}_n$ is an estimate of P in $\mathbf{P}$. In this light, the bootstrap estimate $J_n(\hat{P}_n)$ is a simple *plug-in* estimate of $J_n(P)$.

In nonparametric problems, $\hat{P}_n$ is typically taken to be the empirical distribution of the data. In parametric problems where $\mathbf{P} = \{P_\psi : \psi \in \Psi\}$, $\hat{P}_n$ may be taken to be $P_{\hat{\psi}_n}$, where $\hat{\psi}_n$ is an estimate of ψ.

In general, $J_n(x, \hat{P}_n)$ need not be continuous and strictly increasing in x, so that unique and well-defined quantiles may not exist. To get around this and in analogy to (11.19), define

$$J_n^{-1}(1-\alpha, P) = \inf\{x : J_n(x, P) \geq 1-\alpha\} .$$

If $J_n(\cdot, P)$ has a unique quantile $J_n^{-1}(1-\alpha, P)$, then

$$P\{R_n(X^n, \theta(P)) \leq J_n^{-1}(1-\alpha, P)\} = 1-\alpha ;$$

in general, the probability on the left is at least $1-\alpha$. If $J_n^{-1}(1-\alpha, P)$ were known, then the region

$$\{\theta \in \Theta : \; R_n(X^n, \theta) \leq J_n^{-1}(1-\alpha, P)\}$$

would be a level $1-\alpha$ confidence region for $\theta(P)$. The bootstrap simply replaces $J_n^{-1}(1-\alpha, P)$ by $J_n^{-1}(1-\alpha, \hat{P}_n)$. The resulting bootstrap confidence region for $\theta(P)$ of nominal level $1-\alpha$ takes the form

$$B_n(1-\alpha, X^n) = \{\theta \in \Theta : R_n(X^n, \theta) \leq J_n^{-1}(1-\alpha, \hat{P}_n)\} . \tag{18.3}$$

Suppose the problem is to construct a confidence interval for a real-valued parameter $\theta(P)$ based on the root $|\hat{\theta}_n - \theta(P)|$ for some estimator $\hat{\theta}_n$. The interval (18.3) would then be symmetric about $\hat{\theta}_n$. An alternative equi-tailed interval can be based on the root $\hat{\theta}_n - \theta(P)$ and uses both tails of $J_n(\hat{P}_n)$; it is given by

$$\{\theta \in \Theta : J_n^{-1}(\frac{\alpha}{2}, \hat{P}_n) \leq R_n(X^n, \theta) \leq J_n^{-1}(1-\frac{\alpha}{2}, \hat{P}_n)\} .$$

A comparison of the two approaches will be made in Section 18.4.

Outside certain exceptional cases, the bootstrap approximation $J_n(x, \hat{P}_n)$ cannot be calculated exactly. Even in the relatively simple case when $\theta(P)$ is the mean of P, the root is $n^{1/2}[\bar{X}_n - \theta(P)]$, and $\hat{P}_n$ is the empirical distribution, the exact computation of the bootstrap distribution involves an n-fold convolution.[1] Typically, one resorts to a Monte Carlo approximation to $J_n(P)$, as introduced in Example 11.4.3. Specifically, conditional on the data X^n, for $j = 1, \ldots, B$, let $X_j^{n*} = (X_{1,j}^*, \ldots, X_{n,j}^*)$ be a sample of n i.i.d. observations from $\hat{P}_n$; X_j^{n*} is referred to as the jth bootstrap sample of size n. Of course, when $\hat{P}_n$ is the empirical distribution, this amounts to resampling the original observations with replacement. The bootstrap estimator $J_n(\hat{P}_n)$ is then approximated by the empirical distribution of the B values $R_n(X_j^{n*}, \hat{\theta}_n)$. Because B can be taken to be large (assuming enough computing power), the resulting approximation can be made arbitrarily close to $J_n(\hat{P}_n)$ (see Example 11.4.3), and so we will subsequently focus on the exact bootstrap

[1] Diaconis and Holmes (1994) show how the exact bootstrap distribution can be calculated in some examples.

estimator $J_n(\hat{P}_n)$ while keeping in mind it is usually only approximated by Monte Carlo simulation.

The bootstrap can then be viewed as a simple plug-in estimator of a distribution function. This simple idea, combined with Monte Carlo simulation, allows for quite a broad range of applications.

We will now discuss the consistency of the bootstrap estimator $J_n(\hat{P}_n)$ of the true sampling distribution $J_n(P)$ of $R_n(X^n, \theta(P))$. Typically, one can show that $J_n(P)$ converges weakly to a nondegenerate limit law $J(P)$. Since the bootstrap replaces P by $\hat{P}_n$ in $J_n(\cdot)$, it is useful to study $J_n(P_n)$ under more general sequences $\{P_n\}$. In order to understand the behavior of the random sequence of distributions $J_n(\hat{P}_n)$, it will be easier to first understand how $J_n(P_n)$ behaves for certain fixed sequences $\{P_n\}$. For the bootstrap to be consistent, $J_n(P)$ must be smooth in P since we are replacing P by $\hat{P}_n$. Thus, we are led to studying the asymptotic behavior of $J_n(P_n)$ under fixed sequences of probabilities $\{P_n\}$ which are "converging" to P in a certain sense. Once it is understood how $J_n(P_n)$ behaves for fixed sequences $\{P_n\}$, it is easy to pass to random sequences $\{\hat{P}_n\}$.

In the theorem below, the existence of a continuous limiting distribution is assumed, though its exact form need not be explicit. Although the conditions of the theorem appear strong, they can be verified in many interesting examples.

Theorem 18.3.1 *Let $\mathbf{C}_P$ be a set of sequences $\{P_n \in \mathbf{P}\}$ containing the sequence $\{P, P, \cdots\}$. Suppose that, for every sequence $\{P_n\}$ in $\mathbf{C}_P$, $J_n(P_n)$ converges weakly to a common continuous limit law $J(P)$ having distribution function $J(x, P)$. Let X^n be a sample of size n from P. Assume that $\hat{P}_n$ is an estimate of P based on X^n such that $\{\hat{P}_n\}$ falls in $\mathbf{C}_P$ with probability one. Then,*

$$\sup_x |J_n(x, P) - J_n(x, \hat{P}_n)| \to 0 \textit{ with probability one.} \tag{18.4}$$

If $J(\cdot, P)$ is continuous and strictly increasing at $J^{-1}(1-\alpha, P)$, then

$$J_n^{-1}(1-\alpha, \hat{P}_n) \to J^{-1}(1-\alpha, P) \textit{ with probability one.} \tag{18.5}$$

Also, the bootstrap confidence set $B_n(1-\alpha, X^n)$ given by Eq. (18.3) is pointwise consistent in level; that is,

$$P\{\theta(P) \in B_n(1-\alpha, X^n)\} \to 1-\alpha\ . \tag{18.6}$$

Proof. For the proof of part (18.4), note that the assumptions and Polya's Theorem (Theorem 11.2.9) imply that

$$\sup_x |J_n(x, P) - J_n(x, P_n)| \to 0$$

for any sequence $\{P_n\}$ in $\mathbf{C}_P$. Thus, since $\{\hat{P}_n\} \in \mathbf{C}_P$ with probability one, (18.4) follows. Lemma 11.2.1 implies $J_n^{-1}(1-\alpha, P_n) \to J^{-1}(1-\alpha, P)$ whenever $\{P_n\} \in$

$\mathbf{C}_P$; so (18.5) follows. In order to deduce (18.6), the probability on the left side of (18.6) is equal to

$$P\{R_n(X^n, \theta(P)) \le J_n^{-1}(1-\alpha, \hat{P}_n)\} \,. \tag{18.7}$$

Under P, $R_n(X^n, \theta(P))$ has a limiting distribution $J(\cdot, P)$ and, by (18.5), $J_n^{-1}(1-\alpha, \hat{P}_n) \to J^{-1}(1-\alpha, P)$ with probability one. Thus, by Slutsky's Theorem, (18.7) tends to $J(J^{-1}(1-\alpha, P), P) = 1-\alpha$. ■

Often, the set of sequences $\mathbf{C}_P$ can be described as the set of sequences $\{P_n\}$ such that $d(P_n, P) \to 0$, where d is an appropriate metric on the space of probabilities. Indeed, one should think of $\mathbf{C}_P$ as a set of sequences $\{P_n\}$ that are converging to P in an appropriate sense. Thus, the convergence of $J_n(P_n)$ to $J(P)$ is locally uniform in the sense $d(P_n, P) \to 0$ implies $J_n(P_n)$ converges weakly to $J(P)$. Note, however, that the appropriate metric d will depend on the precise nature of the root.

When the convergences (18.4) and (18.5) hold with probability one, we say the bootstrap is strongly consistent. If these convergences hold in probability, we say the bootstrap is weakly consistent. In any case, (18.6) holds even if (18.4) and (18.5) only hold in probability; see Problem 18.3. Furthermore, the conclusion (18.6) holds if $J(\cdot, P)$ is continuous (and not necessarily strictly increasing); see Problem 18.6.

Example 18.3.1 (**Parametric Bootstrap**) Suppose $X^n = (X_1, \ldots, X_n)$ is a sample from a q.m.d. model $\{P_\theta,\ \theta \in \Omega\}$, where $\Omega \subseteq \mathbb{R}^k$. Suppose $\hat{\theta}_n$ is an efficient likelihood estimator in the sense that (14.62) holds. Let $g(\theta)$ be a differentiable map from Ω to $\mathbb{R}$ with nonzero gradient vector $\dot{g}(\theta)$. Consider the root

$$R_n(X^n, \theta) = n^{1/2}[g(\hat{\theta}_n) - g(\theta)] \,,$$

with distribution function $J_n(x, \theta)$. By Theorem 14.4.1,

$$J_n(x, \theta) \to J(x, \theta) \,,$$

where

$$J(x, \theta) = \Phi(x/\sigma_\theta)$$

and

$$\sigma_\theta^2 = \dot{g}(\theta) I^{-1}(\theta) \dot{g}(\theta)^\top \,.$$

One approach to estimating the distribution of $n^{1/2}[g(\hat{\theta}_n) - g(\theta)]$ is to use the normal approximation $N(0, \hat{\sigma}_n^2)$, where $\hat{\sigma}_n^2$ is a consistent estimator of σ_θ^2. For example, if $\dot{g}(\theta)$ and $I(\theta)$ are continuous in θ, then a weakly consistent estimator of σ_θ^2 is

$$\hat{\sigma}_n^2 = \dot{g}(\hat{\theta}_n) I^{-1}(\hat{\theta}_n) \dot{g}(\hat{\theta}_n)^\top \,.$$

In order to calculate $\hat{\sigma}_n^2$, the forms of $\dot{g}(\cdot)$ and $I(\cdot)$ must be known. This approach of using a normal approximation with an estimator of the limiting variance is a special case of asymptotic approximation discussed in Section 18.2.3. Because it may be difficult to calculate a consistent estimator of the limiting variance, and because the resulting approximation may be poor, it is interesting to consider the bootstrap method. A discussion of higher order asymptotic comparisons will be discussed in Section 18.4. For now, we show the bootstrap approximation $J_n(x, \hat{\theta}_n)$ to $J(x, \theta)$ is weakly consistent.

Theorem 18.3.2 *Under the above setup, under θ,*

$$\sup_x |J_n(x, \theta) - J(x, \theta)| \to 0$$

and

$$\sup_x |J_n(x, \hat{\theta}_n) - J_n(x, \theta)| \to 0 \tag{18.8}$$

in probability; therefore, (18.6) holds.

Proof. For purposes of the proof, assume $k = 1$; the general case is left as an exercise. By Theorem 14.4.1, for any sequence θ_n such that $n^{1/2}(\theta_n - \theta) \to h$, $J_n(x, \theta_n) \to J(x, \theta)$. In trying to apply the previous theorem, define $\mathbf{C}_\theta$ as the set of sequences $\{\theta_n\}$ satisfying $n^{1/2}(\theta_n - \theta) \to h$, for some finite h. (Rather than describe $\mathbf{C}_P$ as a set of sequences of distributions, we identify P_θ with θ and describe $\mathbf{C}_\theta$ as a set of sequences of parameter values.) Unfortunately, $\hat{\theta}_n$ does not fall in $\mathbf{C}_\theta$ with probability one because $n^{1/2}(\hat{\theta}_n - \theta)$ need not converge with probability one. However, we can modify the argument as follows. Since $n^{1/2}(\hat{\theta}_n - \theta)$ converges in distribution, we can apply the Almost Sure Representation Theorem (Theorem 11.4.4). Thus, there exist random variables $\tilde{\theta}_n$ and H defined on a common probability space such that $\hat{\theta}_n$ and $\tilde{\theta}_n$ have the same distribution and $n^{1/2}(\tilde{\theta}_n - \theta) \to H$ almost surely. Then, $\{\tilde{\theta}_n\} \in \mathbf{C}_\theta$ with probability one, and we can conclude

$$\sup_x |J_n(x, \tilde{\theta}_n) - J_n(x, \theta)| \to 0$$

almost surely. Since $\hat{\theta}_n$ and $\tilde{\theta}_n$ have the same distributional properties, so do $J_n(\hat{\theta}_n)$ and $J_n(\tilde{\theta}_n)$, and the result (18.8) follows. ■

A one-sided bootstrap lower confidence bound for $g(\theta)$ takes the form

$$g(\hat{\theta}_n) - n^{-1/2} J_n^{-1}(1 - \alpha, \hat{\theta}_n) \ .$$

The previous theorem implies, under θ,

$$J_n^{-1}(1 - \alpha, \hat{\theta}_n) \xrightarrow{P} \sigma_\theta z_{1-\alpha} \ .$$

Suppose now the problem is to test $g(\theta) = 0$ versus $g(\theta) > 0$. By the duality between tests and confidence regions, one possibility is to reject the null hypothesis if the lower confidence bound exceeds zero, or equivalently when $n^{1/2} g(\hat{\theta}_n) > J_n^{-1}(1 -$

$\alpha, \hat{\theta}_n)$. This test is pointwise asymptotically level α because, by Slutsky's Theorem, $n^{1/2}g(\hat{\theta}_n)$ is asymptotically $N(0, \sigma_\theta^2)$ if $g(\theta) = 0$. The limiting power of this test against a contiguous sequence of alternatives is given in the following corollary.

Corollary 18.3.1 *Under the setup of Example 18.3.1 with θ satisfying $g(\theta) = 0$, the limiting power of the test that rejects when $n^{1/2}g(\hat{\theta}_n) > J_n^{-1}(1-\alpha, \hat{\theta}_n)$ against the sequence $\theta_n = \theta + hn^{-1/2}$ satisfies*

$$P^n_{\theta_n}\{n^{1/2}g(\hat{\theta}_n) > J_n^{-1}(1-\alpha, \hat{\theta}_n)\} \to 1 - \Phi(z_{1-\alpha} - \sigma_\theta^{-1}\langle \dot{g}(\theta)^\top, h\rangle) . \tag{18.9}$$

PROOF. The left-hand side can be written as

$$P^n_{\theta_n}\{n^{1/2}[g(\hat{\theta}_n) - g(\theta_n)] > J_n^{-1}(1-\alpha, \hat{\theta}_n) - n^{1/2}g(\theta_n)\} . \tag{18.10}$$

Under P^n_θ, $J_n^{-1}(1-\alpha, \hat{\theta}_n)$ converges in probability to $\sigma_\theta z_{1-\alpha}$; by contiguity, under $P^n_{\theta_n}$, $J_n^{-1}(1-\alpha, \hat{\theta}_n)$ converges to the same constant. Also, by differentiability of g and the fact that $g(\theta) = 0$

$$n^{1/2}g(\theta_n) \to \langle \dot{g}(\theta)^\top, h\rangle .$$

By Theorem 14.4.1, the left-hand side of (18.10) is asymptotically $N(0, \sigma_\theta^2)$. Letting Z denote a standard normal variable, by Slutsky's Theorem, (18.10) converges to

$$P\{\sigma_\theta Z > \sigma_\theta z_{1-\alpha} - \langle \dot{g}(\theta)^\top, h\rangle\} ,$$

and the result follows. ■

In fact, it follows from Theorem 15.5.1 that this limiting power is optimal. The moral is that the bootstrap can produce an asymptotically optimal test, but only if the initial estimator or test statistic is optimally chosen. Otherwise, if the root is based on a suboptimal estimator, the bootstrap approach to approximating the sampling distribution of a root is so good that the bootstrap will not be optimal. For example, in a normal location model $N(\theta, 1)$, the bootstrap distribution based on the root $\bar{X}_n - \theta$ is exact as previously discussed (except possibly for simulation error), as is the bootstrap distribution for $T_n - \theta$, where T_n is any location equivariant estimator. But, taking T_n equal to the sample median would not lead to an AUMP test, since the bootstrap is approximating the distribution of the sample median, a suboptimal statistic in this case. Furthermore, this leads to the observation that the bootstrap can be used adaptively to approximate several distributions, and then inference can be based on the one with better properties; see Legér and Romano (1990a; 1990b).

In general, one may base the choice of root by an initial estimator $\hat{\theta}_n$ of θ, and then bootstrap the root using $J_n(\tilde{\theta}_n)$, where $\hat{\theta}_n$ and $\tilde{\theta}_n$ may differ. In some instances, the choice is important. For the problem of construction of confidence sets for a multivariate normal mean vector based on the James–Stein estimator, Beran (1995) shows the importance of proper choice of parametric bootstrap, at least when the dimension is moderately high.

18.3.2 The Nonparametric Mean

In this section, we consider the case of Example 18.2.4, confidence intervals for the nonparametric mean. This example deserves special attention because many statistics can be approximated by linear statistics. We will examine this case in detail, since similar considerations apply to more complicated situations. Given a sample $X^n = (X_1, \ldots, X_n)$ from a distribution F on the real line, consider the problem of constructing a confidence interval for $\theta(F) = E_F(X_i)$. Let $\sigma^2(F)$ denote the variance of F. The conditions for Theorem 18.3.1 are verified in the following result.

Theorem 18.3.3 *Let F be a distribution on the line with finite, nonzero variance $\sigma^2(F)$. Let $J_n(F)$ be the distribution of the root $R_n(X^n, \theta(F)) = n^{1/2}[\bar{X}_n - \theta(F)]$.*

(i) Let $\mathbf{C}_F$ be the set of sequences $\{F_n\}$ such that F_n converges weakly to F, $\theta(F_n) \to \theta(F)$, and $\sigma^2(F_n) \to \sigma^2(F)$. If $\{F_n\} \in \mathbf{C}_F$, then $J_n(F_n)$ converges weakly to $J(F)$, where $J(F)$ is the normal distribution with mean zero and variance $\sigma^2(F)$.

(ii) Let $X_1, \ldots, X_n$ be i.i.d. F, and let $\hat{F}_n$ denote the empirical distribution function. Then, the bootstrap estimator $J_n(\hat{F}_n)$ is strongly consistent so that (18.4), (18.5), and (18.6) hold.

Proof of Theorem 18.3.3. For the purpose of proving (i), construct variables $X_{n,1}, \ldots, X_{n,n}$ which are independent with identical distribution F_n, and set $\bar{X}_n = \sum_i X_{n,i}/n$. We must show that the law of $n^{1/2}(\bar{X}_n - \mu(F_n))$ converges weakly to $J(F)$. It suffices to verify the Lindeberg Condition for $Y_{n,i}$, where $Y_{n,i} = X_{n,i} - \mu(F_n)$. This entails showing that, for each $\epsilon > 0$,

$$\lim_{n \to \infty} E[Y_{n,1}^2 1(Y_{n,1}^2 > n\epsilon^2)] = 0 . \tag{18.11}$$

Note that $Y_{n,1} \xrightarrow{d} Y$, where $Y = X - \mu(F)$ and X has distribution F, and $E(Y_{n,1}^2) \to E(Y^2)$. By the continuous mapping theorem (Theorem 11.2.10), $Y_{n,1}^2 \xrightarrow{d} Y^2$. Now, for any fixed $\beta > 0$ and all $n > \beta/\epsilon^2$,

$$E[Y_{n,1}^2 1(Y_{n,1}^2 > n\epsilon^2)] \le E[Y_{n,1}^2 1(Y_{n,1}^2 > \beta)] \to E[Y^2 1(Y^2 > \beta)] ,$$

where the last convergence holds if β is a continuity point of the distribution of Y^2, by (11.40). Since the set of continuity points of any distribution is dense and $E[Y^2 1(Y^2 > \beta)] \downarrow 0$ as $\beta \to \infty$, Lindeberg's Condition holds.

We now prove (ii) by applying Theorem 18.3.1; we must show that $\{\hat{F}_n\} \in \mathbf{C}_F$ with probability one. By the Glivenko–Cantelli Theorem,

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0 \quad \text{with probability one} .$$

Also, by the Strong Law of Large Numbers, $\theta(\hat{F}_n) \to \theta(F)$ with probability one and $\sigma^2(\hat{F}_n) \to \sigma^2(F)$ with probability one. Thus, bootstrap confidence intervals for the mean based on the root $R_n(X^n, \theta(F)) = n^{1/2}(\bar{X}_n - \theta(F))$ are asymptotically consistent in the sense of the theorem. ■

Remark 18.3.1 Let F and G be two distribution functions on the real line and define $d_p(F, G)$ to be the infimum of $\{E[|X - Y|^p]\}^{1/p}$ over all pairs of random variables X and Y such that X has distribution F and Y has distribution G. It can be shown that the infimum is attained and that d_p is a metric on the space of distributions having a pth moment. Further, if F has a finite variance $\sigma^2(F)$, then $d_2(F_n, F) \to 0$ is equivalent to F_n converging weakly to F and $\sigma^2(F_n) \to \sigma^2(F)$. Hence, Theorem 18.3.3 may be restated as follows. If F has a finite variance $\sigma^2(F)$ and $d_2(F_n, F) \to 0$, then $J_n(F_n)$ converges weakly to $J(F)$. The metric d_2 is known as Mallow's metric. For details, see Bickel and Freedman (1981).

Continuing the example of the nonparametric mean, it is of interest to consider roots other than $n^{1/2}(\bar{X}_n - \theta(F))$. Specifically, consider the studentized root

$$R_n^s(X^n, \theta(F)) = n^{1/2}(\bar{X}_n - \theta(F))/\sigma(\hat{F}_n) , \tag{18.12}$$

where $\sigma^2(\hat{F}_n)$ is the usual bootstrap estimate of variance. To obtain consistency of the bootstrap method, called the bootstrap-t, we appeal to the following result.

Theorem 18.3.4 *Suppose F is a c.d.f. with finite nonzero variance $\sigma^2(F)$. Let $K_n(F)$ be the distribution of the root (18.12) based on a sample of size n from F.*

(i) Let $\mathbf{C}_F$ be defined as in Theorem 18.3.3. Then, for any sequence $\{F_n\} \in \mathbf{C}_F$, $K_n(F_n)$ converges weakly to the standard normal distribution.

(ii) Hence, the bootstrap sampling distribution $K_n(\hat{F}_n)$ is consistent in the sense that (18.4), (18.5), and (18.6) hold.

Before proving this theorem, we first need a weak law of large numbers for a triangular array that generalizes Theorem 11.3.1. The following lemma serves as a suitable version for our purposes.

Lemma 18.3.1 *Suppose $Y_{n,1}, \ldots, Y_{n,n}$ is a triangular array of independent random variables, the n-th row having c.d.f. G_n. Assume G_n converges in distribution to G and*

$$E[|Y_{n,1}|] \to E[|Y|] < \infty$$

as $n \to \infty$, where Y has c.d.f. G. Then,

$$\bar{Y}_n \equiv n^{-1} \sum_{i=1}^{n} Y_{n,i} \xrightarrow{P} E(Y)$$

as $n \to \infty$.

Proof. Apply Lemma 13.4.2 and (11.40). ■

PROOF OF THEOREM 18.3.4. For the proof, let $X_{n,1}, \ldots, X_{n,n}$ be independent with distribution F_n. By Theorem 18.3.3 and Slutsky's Theorem, it is enough to show $\sigma^2(\hat{F}_n) \to \sigma^2(F)$ in probability under F_n. But,

$$\sigma^2(\hat{F}_n) = \frac{1}{n} \sum_i (X_{n,i} - \bar{X}_n)^2 .$$

Now, apply Lemma 18.3.1 on the Weak Law of Large Numbers for a triangular array with $Y_{n,i} = X_{n,i}$ and also with $Y_{n,i} = X_{n,i}^2$. The consistency of the bootstrap method based on the root (18.12) now follows easily. ■

It is interesting to consider how the bootstrap behaves when the underlying distribution has an infinite variance (but well-defined mean). The short answer is that the bootstrap procedure considered thus far will fail, in the sense that the convergence in expression (18.4) does not hold. The failure of the bootstrap for the mean in the infinite variance case was first noted by Babu (1984); further elucidation is given in Athreya (1987) and Knight (1989). In fact, a striking theorem due to Giné and Zinn (1989) asserts that the simple bootstrap studied thus far will work for the mean in the sense of strong consistency if and only if the variance is finite. For a nice exposition of related results, see Giné (1997).

Related results for the studentized bootstrap based on approximating the distribution of the root (18.12) were considered by Csörgö and Mason (1989) and Hall (1990). The conclusion is that the bootstrap is strongly or almost surely consistent if and only if the variance is finite; the bootstrap is weakly consistent if and only if X_i is in the domain of attraction of the normal distribution.

In fact, it was realized by Athreya (1985) that the bootstrap can be modified so that consistency ensues even with infinite variance. The modification consists of reducing the bootstrap sample size. Further results are given in Arcones and Giné (1989, 1991). In other instances where the simple bootstrap fails, consistency can often be recovered by reducing the bootstrap sample size. The benefit of reducing the bootstrap sample size was recognized first in Bretagnolle (1983). An even more general approach based on subsampling will be considered later in Section 18.7.

18.3.3 Further Examples

Example 18.3.2 (**Multivariate Mean**) Let $X^n = (X_1, \ldots, X_n)$ be a sample of n observations from F, where X_i takes values in $\mathbb{R}$. Let $\theta(F) = E_F(X_i)$ be equal to the mean vector, and let

$$S_n(X^n, \theta(F)) = n^{1/2}(\bar{X}_n - \theta(F)) , \tag{18.13}$$

where $\bar{X}_n = \sum_i X_i/n$ is the sample mean vector. Let

$$R_n(X^n, \theta(F)) = \left\| S_n(X^n, \theta(F)) \right\| ,$$

where $\| \cdot \|$ is any norm on $\mathbb{R}^k$. The consistency of the bootstrap method based on the root R_n follows from the following theorem.

Theorem 18.3.5 *Let $L_n(F)$ be the distribution (in $\mathbb{R}^k$) of $S_n(X^n, \theta(F))$ under F, where S_n is defined in (18.13). Let $\Sigma(F)$ be the covariance matrix of S_n under F. Let $\mathbf{C}_F$ be the set of sequences $\{F_n\}$ such that F_n converges weakly to F and $\Sigma(F_n) \to \Sigma(F)$, so that each entry of the matrix $\Sigma(F_n)$ converges to the corresponding entry (assumed finite) of $\Sigma(F)$.*

(i) Then, $L_n(F_n)$ converges weakly to $L(F)$, the multivariate normal distribution with mean zero and covariance matrix $\Sigma(F)$.

(ii) Assume $\Sigma(F)$ contains at least one nonzero component. Let $\| \cdot \|$ be any norm on $\mathbb{R}$ and let $J_n(F)$ be the distribution of $R_n(X^n, \theta(F)) = \|S_n(X^n, \theta(F))\|$ under F. Then, $J_n(F_n)$ converges weakly to $J(F)$, which is the distribution of $\|Z\|$ when Z has distribution $L(F)$.

(iii) Suppose $X_1, \ldots, X_n$ are i.i.d. F with empirical distribution $\hat{F}_n$ (in $\mathbb{R}$). Then, the bootstrap approximation satisfies

$$\rho(J_n(F), J_n(\hat{F}_n)) \to 0 \textit{ with probability one} ,$$

and bootstrap confidence regions based on the root R_n are consistent in the sense that the convergences (18.4) to (18.6) hold.

Proof. The proof of (i) follows by the Cramer–Wold device (Theorem 11.2.3) and by Theorem 18.3.3 (i). To prove (ii), note that any norm $\| \cdot \|$ on $\mathbb{R}$ is continuous almost everywhere with respect to $L(F)$. A proof of this statement can be based on the fact that, for any norm $\| \cdot \|$, the set $\{x \in \mathbb{R} : \|x\| = c\}$ has Lebesgue measure zero because it is the boundary of a convex set. So, the continuous mapping theorem applies and so $J_n(F_n)$ converges weakly to $J(F)$.

Part (iii) follows because $\{\hat{F}_n\} \in \mathbf{C}_F$ with probability one, by the Glivenko–Cantelli Theorem (on $\mathbb{R}$) and the strong law of large numbers. ■

Note the power of the bootstrap method. Analytical methods for approximating the distribution of the root $R_n = \|S_n\|$ would depend heavily on the choice of norm $\| \cdot \|$, but the bootstrap handles them all with equal ease.

Let $\hat{\Sigma}_n = \Sigma(\hat{F})$ be the sample covariance matrix. As in the univariate case, one can also bootstrap the root defined by

$$\tilde{R}_n(X^n, \theta(F)) = \|\hat{\Sigma}_n^{-1/2}(\bar{X}_n - \theta(F))\|, \tag{18.14}$$

provided $\Sigma(F)$ is assumed positive definite. In the case where $\| \cdot \|$ is the usual Euclidean norm, this root leads to confidence ellipsoid, i.e., a confidence set whose shape is an ellipsoid.

Example 18.3.3 (**Smooth Functions of Means**) Let $X_1, \ldots, X_n$ be i.i.d. S-valued random variables with distribution P. Suppose $\theta = \theta(P) = (\theta_1, \ldots, \theta_p)$, where $\theta_j = E_P[h_j(X_i)]$ and the h_j are real-valued functions defined on S. Interest focuses on θ or some function f of θ. Let $\hat{\theta}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,p})$, where $\hat{\theta}_{n,j} = \sum_{i=1}^n h_j(X_i)/n$. Assume moment conditions on the $h_j(X_i)$. Then, by the multivariate mean case, the bootstrap approximation to the distribution of $n^{1/2}(\hat{\theta}_n - \theta)$ is appropriately close in the sense

$$\rho\left(\mathcal{L}_P(n^{1/2}(\hat{\theta}_n - \theta)), \mathcal{L}_{P_n^*}(n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n))\right) \to 0 \qquad (18.15)$$

with probability one, where ρ is any metric metrizing weak convergence in $\mathbb{R}^p$ (such as the Bounded–Lipschitz metric introduced in Problem 11.24). Here, P_n^* refers to the distribution of the data resampled from the empirical distribution conditional on $X_1, \ldots X_n$. Moreover,

$$\rho\left(\mathcal{L}_P(n^{1/2}(\hat{\theta}_n - \theta)), \mathcal{L}(Z)\right) \to 0 \, , \qquad (18.16)$$

where Z is multivariate normal with mean zero and covariance matrix Σ having (i, j)-th component

$$Cov(Z_i, Z_j) = Cov[h_i(X_1), h_j(X_1)].$$

To see why, define Y_i to be the vector in $\mathbb{R}^p$ with j-th component $h_j(X_i)$, so that we are exactly back in the multivariate mean case. Now, suppose f is an appropriately smooth function from $\mathbb{R}^p$ to $\mathbb{R}^q$, and interest now focuses on the parameter $\mu = f(\theta)$. Assume $f = (f_1, \ldots, f_q)^\top$, where $f_i(y_1, \ldots, y_p)$ is a real-valued function from $\mathbb{R}^p$ having a nonzero continuous differential at $(y_1, \cdots, y_p) = (\theta_1, \ldots, \theta_p)$. Let D be the $q \times p$ matrix with (i, j) entry $\partial f_i(y_1, \ldots, y_p)/\partial y_j$ evaluated at $(\theta_1, \ldots, \theta_p)$. Then, the following is true.

Theorem 18.3.6 *Suppose f is a function satisfying the above smoothness assumptions. If $E[h_j^2(X_i)] < \infty$, then Eqs. (18.15) and (18.16) hold. Moreover,*

$$\rho\left(\mathcal{L}_P(n^{1/2}[f(\hat{\theta}_n) - f(\theta)]), \mathcal{L}_{P_n^*}(n^{1/2}[f(\hat{\theta}_n^*) - f(\hat{\theta}_n)])\right) \to 0$$

with probability one and

$$\sup_s \left| P\{\|f(\hat{\theta}_n) - f(\theta)\| \le s\} - P_n^*\{\|f(\hat{\theta}_n^*) - f(\hat{\theta}_n)\| \le s\} \right| \to 0$$

with probability one.

Proof. The proof follows as Eqs. (18.15) and (18.16) are immediate from the multivariate mean case. The smoothness assumptions on f and the Delta Method imply that $n^{1/2}[f(\hat{\theta}_n) - f(\theta)]$ has a limiting multivariate normal distribution with mean 0 and covariance matrix $D\Sigma D^\top$; see Theorem 11.3.4. Similar arguments apply to the bootstrap counterpart. Details are left to the reader (Problem 18.18). ■

Example 18.3.4 (**Joint Confidence Rectangles**) Under the assumptions of Theorem 18.3.6, a joint confidence set can be constructed for $(f_1(\theta), \ldots, f_q(\theta))$ with asymptotic coverage $1 - \alpha$. In the case where $\|x\| = \max |x_i|$, the set is a rectangle in $\mathbb{R}^q$. Such a set is easily described as

$$\{f(\theta) : \ |f_i(\hat{\theta}_n) - f_i(\theta)| \le \hat{b}_n(1-\alpha) \quad \text{for all } i \ \},$$

where $\hat{b}_n(1-\alpha)$ is the bootstrap approximation to the $1-\alpha$ quantile of the distribution of $\max_i |f_i(\hat{\theta}_n) - f_i(\theta)|$. Thus, a value for $f_i(\theta)$ is included in the region if and only if $f_i(\theta) \in f_i(\hat{\theta}_n) \pm \hat{b}_n(1-\alpha)$. Note, however, the intervals $f_i(\hat{\theta}_n) \pm \hat{b}_n(1-\alpha)$ may be unbalanced in the sense that the limiting coverage probability for each marginal parameter $f_i(\theta)$ may depend on i. To fix this, one could instead bootstrap the distribution of $\max_i |f_i(\hat{\theta}_n) - f_i(\theta)|/\hat{\sigma}_{n,i}$, where $\hat{\sigma}^2_{n,i}$ is some consistent estimate of the (i, i) entry of the asymptotic covariance matrix $D\Sigma D^\top$ for $n^{1/2} f(\hat{\theta}_n)$. For further discussion, see Beran (1988a), who employs a transformation called prepivoting to achieve balance. ■

Example 18.3.5 (**Uniform Confidence Bands for a c.d.f.** $\boldsymbol{F}$) Consider a sample $X^n = (X_1, \ldots, X_n)$ of real-valued observations having c.d.f. F. The empirical c.d.f. $\hat{F}_n$ is then

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} I\{X_i \le t\} .$$

For two distribution functions F and G, define the Kolmogorov–Smirnov (or uniform) metric

$$d_K(F, G) = \sup_t |F(t) - G(t)| .$$

Now, consider the root

$$R_n(X^n, \theta(F)) = n^{1/2} d_K(\hat{F}_n, F) ,$$

whose distribution under F is denoted by $J_n(F)$. As discussed in Example 11.4.2, $J_n(F)$ has a continuous limiting distribution. In fact, the following triangular array convergence holds. If $d_K(F_n, F) \to 0$, then $J_n(F_n) \xrightarrow{d} J(F)$; for a proof, see Politis et al. (1999, p. 20). Thus, we can define $\mathbf{C}_F$ to be the set of sequences $\{F_n\}$ satisfying $d_K(F_n, F) \to 0$. By the Glivenko–Cantelli Theorem, $d_K(\hat{F}_n, F) \to 0$ with probability one, and strong consistency of the bootstrap follows. The resulting uniform confidence bands for F are then consistent in the sense that (18.6) holds, and no assumption on continuity of F is needed (unlike the classical limit theory). This example has been generalized considerably, and the proof depends on the behavior of $n^{1/2}[\hat{F}_n(t) - F(t)]$, which can be viewed as a random function and is called the *empirical process*. The general theory of bootstrapping the empirical processes is developed in van der Vaart and Wellner (1996) and in Chapter 2 of Giné (1997). In particular, the theory generalizes to quite general spaces S, so that the observations need not be real-valued. In the special case when S is k-dimensional Euclidean space, the k-dimensional empirical process was considered in Beran and Millar (1986). Confidence sets for a multivariate distribution based on the bootstrap can then be constructed which are pointwise consistent in level. ■

18.4 Higher Order Asymptotic Comparisons

One of the main reasons the bootstrap approach is so valuable is that it can be applied to approximate the sampling distribution of an estimator in situations where the finite-sample or large-sample distribution theory is intractable, or depends on unknown parameters. However, even in relatively simple situations, we will see that there are advantages to using a bootstrap approach. For example, consider the problem of constructing a confidence interval for a mean. Under the assumption of a finite variance, the standard normal theory interval and the bootstrap-t are each pointwise consistent in level. In order to compare them, we must consider higher order asymptotic properties. More generally, suppose I_n is a nominal $1 - \alpha$ level confidence interval for a parameter $\theta(P)$. Its coverage error under P is

$$P\{\theta(P) \in I_n\} - (1 - \alpha) \ ,$$

and we would like to examine the rate at which this tends to zero. In typical problems, this coverage error is a power of $n^{-1/2}$. It will be necessary to distinguish one-sided and two-sided confidence intervals because their orders of coverage error may differ.

Throughout this section, attention will focus on confidence intervals for the mean in a nonparametric setting. Specifically, we would like to compare some asymptotic methods based on the normal approximation and the bootstrap. Let $X^n = (X_1, \ldots, X_n)$ be i.i.d. with c.d.f. F, mean $\theta(F)$, and variance $\sigma^2(F)$. Also, let $\hat{F}_n$ denote the empirical c.d.f., and let $\hat{\sigma}_n = \sigma(\hat{F}_n)$.

Before addressing coverage error, we recall from Section 13.3 the Edgeworth expansions for the distributions of the roots

$$R_n(X^n, F) = n^{1/2}(\bar{X}_n - \theta(F))$$

and

$$R_n^s(X^n, F) = n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n \ ;$$

as in Section 18.3.2, their distribution functions under F are denoted by $J_n(\cdot, F)$ and $K_n(\cdot, F)$, respectively. Let Φ and φ denote the standard normal c.d.f. and density, respectively.

Theorem 18.4.1 *Assume $E_F(X_i^4) < \infty$. Let ψ_F denote the characteristic function of F, and assume*

$$\limsup_{|s| \to \infty} |\psi_F(s)| < 1 . \tag{18.17}$$

Then,

$$J_n(t, F) = \Phi(t/\sigma(F)) - \frac{1}{6}\gamma(F)\varphi(t/\sigma(F))(\frac{t^2}{\sigma^2(F)} - 1)n^{-1/2} + O(n^{-1}) , \tag{18.18}$$

where

$$\gamma(F) = E_F[X_1 - \theta(F)]^3/\sigma^3(F)$$

is the skewness of F. Moreover, the expansion holds uniformly in t in the sense that

$$J_n(t, F) = [\Phi(t/\sigma(F)) - \frac{1}{6}\gamma(F)\varphi(t/\sigma(F))(\frac{t^2}{\sigma^2(F)} - 1)n^{-1/2}] + R_n(t, F) ,$$

where $|R_n(t, F)| \le C/n$ for all t and some $C = C_F$ which depends on F.

Theorem 18.4.2 *Assume $E_F(X_i^4) < \infty$ and that F is absolutely continuous. Then, uniformly in t,*

$$K_n(t, F) = \Phi(t) + \frac{1}{6}\gamma(F)\varphi(t)(2t^2 + 1)n^{-1/2} + O(n^{-1}) . \tag{18.19}$$

Note that the term of order $n^{-1/2}$ is zero if and only if the underlying skewness $\gamma(F)$ is zero, so that the dominant error in using a standard normal approximation to the distribution of the studentized statistic is due to skewness of the underlying distribution. We will use these expansions in order to derive some important properties of confidence intervals. Note, however, that the expansions are asymptotic results, and for finite n, including the correction term (i.e., the term of order $n^{-1/2}$) may worsen the approximation.

Expansions for the distribution of a root such as (18.18) and (18.19) imply corresponding expansions for their quantiles, which are known as *Cornish–Fisher Expansions*. For example, $K_n^{-1}(1-\alpha, F)$ is a value of t satisfying $K_n(t, F) = 1 - \alpha$. Of course, $K_n^{-1}(1-\alpha, F) \to z_{1-\alpha}$. We would like to determine $c = c(\alpha, F)$ such that

$$K_n^{-1}(1-\alpha, F) = z_{1-\alpha} + cn^{-1/2} + O(n^{-1}) .$$

Set $1 - \alpha$ equal to the right-hand side of (18.19) with $t = z_{1-\alpha} + cn^{-1/2}$, which yields

$$\Phi(z_{1-\alpha} + cn^{-1/2}) + \frac{1}{6}\gamma(F)\varphi(z_{1-\alpha} + cn^{-1/2})(2z_{1-\alpha}^2 + 1)n^{-1/2} + O(n^{-1}) = 1 - \alpha .$$

By expanding Φ and φ about $z_{1-\alpha}$, we find that

$$c = -\frac{1}{6}\gamma(F)(2z_{1-\alpha}^2 + 1) .$$

Thus,

$$K_n^{-1}(1-\alpha, F) = z_{1-\alpha} - \frac{1}{6}\gamma(F)(2z_{1-\alpha}^2 + 1)n^{-1/2} + O(n^{-1}) . \tag{18.20}$$

In fact, under the assumptions of Theorem 18.4.2, the expansion (18.19) holds uniformly in t, and so the expansion (18.20) holds uniformly in $\alpha \in [\epsilon, 1-\epsilon]$, for any $\epsilon > 0$ (Problem 18.20). Similarly, one can show (Problem 18.21) that, under the assumptions of Theorem 18.4.1,

$$J_n^{-1}(1-\alpha, F) = \sigma(F)z_{1-\alpha} + \frac{1}{6}\sigma(F)\gamma(F)(z_{1-\alpha}^2 - 1)n^{-1/2} + O(n^{-1}) , \tag{18.21}$$

uniformly in $\alpha \in [\epsilon, 1-\epsilon]$.

Normal Theory Intervals. The most basic approximate upper one-sided confidence interval for the mean $\theta(F)$ is given by

$$\bar{X}_n + n^{-1/2}\hat{\sigma}_n z_{1-\alpha} , \tag{18.22}$$

where $\hat{\sigma}_n^2 = \sigma^2(\hat{F}_n)$ is the (biased) sample variance. Its one-sided coverage error is given by

$$P_F\{\theta(F) \le \bar{X}_n + n^{-1/2}\hat{\sigma}_n z_{1-\alpha}\} - (1-\alpha)$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha\} . \tag{18.23}$$

By (18.19), the one-sided coverage error of this normal theory interval is

$$-\frac{1}{6}\gamma(F)\varphi(z_\alpha)(2z_\alpha^2 + 1)n^{-1/2} + O(n^{-1}) = O(n^{-1/2}) . \tag{18.24}$$

Analogously, the coverage error of the two-sided confidence interval of nominal level $1-2\alpha$,

$$\bar{X}_n \pm n^{-1/2}\hat{\sigma}_n z_{1-\alpha} , \tag{18.25}$$

satisfies

$$P_F\{-z_{1-\alpha} \le n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n \le z_{1-\alpha}\} - (1-2\alpha)$$

$$= P\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n \le z_{1-\alpha}\} - P\{n^{1/2}(\bar{X}_n - \theta(F))\hat{\sigma}_n < -z_{1-\alpha}\} - (1-2\alpha) ,$$

which by (18.19) is equal to

$$[\Phi(z_{1-\alpha})+\frac{1}{6}\gamma(F)\varphi(z_{1-\alpha})(2z_{1-\alpha}^2+1)n^{-1/2}+O(n^{-1})]$$

$$-[\Phi(-z_{1-\alpha})+\frac{1}{6}\gamma(F)\varphi(-z_{1-\alpha})(2z_{1-\alpha}^2+1)n^{-1/2}+O(n^{-1})]-(1-2\alpha)=O(n^{-1})\ ,$$

using the symmetry of the function φ. Thus, while the coverage error of the one-sided interval (18.22) is $O(n^{-1/2})$, the two-sided interval (18.25) has coverage error $O(n^{-1})$. The main reason the one-sided interval has coverage error $O(n^{-1/2})$ derives from the fact that a normal approximation is used for the distribution of $n^{1/2}(\bar{X}_n-\theta(F))/\hat{\sigma}_n$ and no correction is made for skewness of the underlying distribution. For example, if $\gamma(F)>0$, the one-sided upper confidence bound (18.22) undercovers slightly while the one-sided lower confidence bound overcovers. The combination of overcoverage and undercoverage yields a net result of a reduction in the order of coverage error of two-sided intervals. Analytically, this fact derives from the key property that the $n^{-1/2}$ term in (18.19) is an even polynomial. (Note, however, that the one-sided coverage error is $O(n^{-1})$ if $\gamma(F)=0$.) These results are in complete analogy with the corresponding results in Section 13.3 for error in rejection probability of tests of the mean based on the normal approximation.

Basic Bootstrap Intervals. Next, we consider bootstrap confidence intervals for $\theta(F)$ based on the root

$$R_n(X^n,\theta(F))=n^{1/2}(\bar{X}_n-\theta(F))\ . \tag{18.26}$$

It is plausible that the bootstrap approximation $J_n(t,\hat{F}_n)$ to $J_n(t,F)$ satisfies an expansion like (18.18) with F replaced by $\hat{F}_n$. In fact, it is the case that

$$J_n(t,\hat{F}_n)=\Phi(t/\hat{\sigma}_n)-\frac{1}{6}\gamma(\hat{F}_n)\varphi(t/\hat{\sigma}_n)(\frac{t^2}{\hat{\sigma}_n^2}-1)n^{-1/2}+O_P(n^{-1})\ . \tag{18.27}$$

Both sides of (18.27) are random and the remainder term is now of order n^{-1} in probability. Similarly, the bootstrap quantile function $J_n^{-1}(1-\alpha,\hat{F}_n)$ has an analogous expansion to (18.21) and is given by

$$J_n^{-1}(1-\alpha,\hat{F}_n)=\hat{\sigma}_n[z_{1-\alpha}+\frac{1}{6}\gamma(\hat{F}_n)(z_{1-\alpha}^2-1)n^{-1/2}]+O_P(n^{-1})\ . \tag{18.28}$$

The validity of these expansions is quite technical and is proved in Hall (1992, Section 5.2), and a sufficient condition for them to hold is that F satisfies Cramér's condition and has infinitely many moments; such assumptions will remain in force for the remainder of this section. From (18.18) and (18.27), it follows that

$$J_n(t,\hat{F}_n)-J_n(t,F)=O_P(n^{-1/2})$$

because

$$\hat{\sigma}_n-\sigma(F)=O_P(n^{-1/2})\ .$$

Thus, the bootstrap approximation $J_n(t, \hat{F}_n)$ to $J_n(t, F)$ has the same order of error as that provided by the normal approximation.

Turning now to coverage error, consider the one-sided coverage error of the nominal level $1-\alpha$ upper confidence bound $\bar{X}_n - n^{-1/2} J_n^{-1}(\alpha, \hat{F}_n)$, given by

$$P_F\{\theta(F) \le \bar{X}_n - n^{-1/2} J_n^{-1}(\alpha, \hat{F}_n)\} - (1-\alpha)$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F)) < J_n^{-1}(\alpha, \hat{F}_n)\}$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha + \frac{1}{6}\gamma(F)(z_\alpha^2 - 1)n^{-1/2} + O_P(n^{-1})\}$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha + \frac{1}{6}\gamma(F)(z_\alpha^2 - 1)n^{-1/2}\} + O(n^{-1})\ .$$

The last equality, though plausible, requires a rigorous argument, but follows from Problem 18.22. The last expression, by (18.19) and a Taylor expansion, becomes

$$-\frac{1}{2}\gamma(F)\varphi(z_\alpha)z_\alpha^2 n^{-1/2} + O(n^{-1})\ ,$$

so that the one-sided coverage error is of the same order as that provided by the basic normal approximation. Moreover, by similar reasoning, the two-sided bootstrap interval of nominal level $1-2\alpha$, given by

$$[\bar{X}_n - n^{-1/2} J_n^{-1}(1-\alpha, \hat{F}_n), \bar{X}_n - n^{-1/2} J_n^{-1}(\alpha, \hat{F}_n)]\ , \tag{18.29}$$

has coverage error $O(n^{-1})$. Although these basic bootstrap intervals have the same orders of coverage error as those based on the normal approximation, there is evidence that the bootstrap does provide some improvement (in terms of the size of the constants); see Liu and Singh (1987).

Bootstrap-t Confidence Intervals. Next, we consider bootstrap confidence intervals for $\theta(F)$ based on the studentized root

$$R_n^s(X^n, \theta(F)) = n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n\ , \tag{18.30}$$

whose distribution under F is denoted by $K_n(\cdot, F)$. The bootstrap versions of the expansions (18.19) and (18.20) are

$$K_n(t, \hat{F}_n) = \Phi(t) + \frac{1}{6}\gamma(\hat{F}_n)\varphi(t)(2t^2+1)n^{-1/2} + O_P(n^{-1}) \tag{18.31}$$

and

$$K_n^{-1}(1-\alpha, \hat{F}_n) = z_{1-\alpha} - \frac{1}{6}\gamma(\hat{F}_n)(2z_{1-\alpha}^2 + 1)n^{-1/2} + O_P(n^{-1})\ . \tag{18.32}$$

Again, these results are obtained rigorously in Hall (1992), and a sufficient condition for their validity is that F is absolutely continuous with infinitely many moments. By comparing (18.19) and (18.31), it follows that

$$K_n(t, \hat{F}_n) - K_n(t, F) = O_P(n^{-1}) \;, \tag{18.33}$$

since $\gamma(\hat{F}_n) - \gamma(F) = O_P(n^{-1/2})$. Similarly,

$$K_n^{-1}(1-\alpha, \hat{F}_n) - K_n^{-1}(1-\alpha, F) = O_P(n^{-1}) \;. \tag{18.34}$$

Thus, the bootstrap is more successful at estimating the distribution or quantiles of the studentized root than its non-studentized version.

Now, consider the nominal level $1-\alpha$ upper confidence bound $\bar{X}_n - n^{-1/2}\hat{\sigma}_n K_n^{-1}(\alpha, \hat{F}_n)$. Its coverage error is given by

$$P_F\{\theta(F) \le \bar{X}_n - n^{-1/2}\hat{\sigma}_n K_n^{-1}(\alpha, \hat{F}_n)\} - (1-\alpha)$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < K_n^{-1}(\alpha, \hat{F}_n)\}$$

$$= \alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha - \frac{1}{6}\gamma(F)(2z_\alpha^2 + 1)n^{-1/2} + O_P(n^{-1})\} \;,$$

since (18.32) implies the same expansion for $K_n^{-1}(\alpha, \hat{F}_n)$ with $\gamma(\hat{F}_n)$ replaced by $\gamma(F)$ (again using the fact that $\gamma(\hat{F}_n) - \gamma(F) = O_P(n^{-1/2})$). By Problem 18.22, this last expression becomes

$$\alpha - P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha - \frac{1}{6}\gamma(F)(2z_\alpha^2 + 1)n^{-1/2}\} + O(n^{-1}) \;.$$

Let

$$t_n = t_n(\alpha, F) = z_\alpha - \frac{1}{6}\gamma(F)(2z_\alpha^2 + 1)n^{-1/2} \;,$$

so that $(t_n - z_\alpha) = O(n^{-1/2})$. Then, the coverage error becomes

$$\alpha - [\Phi(t_n) + \frac{1}{6}\gamma(F)\varphi(t_n)(2t_n^2 + 1)n^{-1/2} + O(n^{-1})] \;.$$

By expanding Φ and φ about z_α and combining terms that are $O(n^{-1})$, the last expression becomes

$$\alpha - \Phi(z_\alpha) - (t_n - z_\alpha)\varphi(z_\alpha) + O(n^{-1})$$

$$-\frac{1}{6}\gamma(F)[\varphi(z_\alpha) + (t_n - z_\alpha)\varphi'(z_\alpha) + O(n^{-1})](2z_\alpha^2 + 1)n^{-1/2} + O(n^{-1}) = O(n^{-1}) \;.$$

Thus, the one-sided coverage error of the bootstrap-t interval is $O(n^{-1})$ and is of smaller order than that provided by the normal approximation or the bootstrap based on a non-studentized root. Intervals with one-sided coverage error of order $O(n^{-1})$ are said to be *second-order accurate*, while intervals with one-sided coverage error of order $O(n^{-1/2})$ are only *first-order accurate*.

A heuristic reason why the bootstrap based on the root (18.30) outperforms the bootstrap based on the root (18.26) is as follows. In the case of (18.26), the bootstrap is estimating a distribution that has mean 0 and unknown variance $\sigma^2(F)$. The main contribution to the estimation error is the implicit estimation of $\sigma^2(F)$ by $\sigma^2(\hat{F}_n)$. On the other hand, the root (18.30) has a distribution that is nearly independent of F since it is an asymptotic pivot.

The two-sided interval of nominal level $1-2\alpha$,

$$[\bar{X}_n - n^{-1/2}\hat{\sigma}_n K_n^{-1}(1-\alpha, \hat{F}_n), \bar{X}_n - n^{-1/2}\hat{\sigma}_n K_n^{-1}(\alpha, \hat{F}_n)] , \qquad (18.35)$$

also has coverage error $O(n^{-1})$ (Problem 18.24). This interval was formed by combining two one-sided intervals. Instead, consider the absolute studentized root

$$R_n^t(X^n, \theta(F)) = |n^{1/2}(\bar{X}_n - \theta(F))|/\hat{\sigma}_n ,$$

whose distribution and quantile functions under F are denoted by $L_n(t, F)$ and $L_n^{-1}(1-\alpha, F)$, respectively. An alternative two-sided bootstrap confidence interval for $\theta(F)$ of nominal level $1-\alpha$ is given by

$$\bar{X}_n \pm n^{-1/2}\hat{\sigma}_n L_n^{-1}(1-\alpha, \hat{F}_n) .$$

Note that this interval is symmetric about $\bar{X}_n$. Its coverage error is actually $O(n^{-2})$. The arguments for this claim are similar to the previous claims about coverage error, but more terms are required in expansions like (18.19).

Bootstrap Calibration. By considering a studentized statistic, the bootstrap-t yields one-sided confidence intervals with coverage error smaller than the non-studentized case. However, except in some simple problems, it may be difficult to standardize or studentize a statistic because an explicit estimate of the asymptotic variance may not be available. An alternative approach to improving coverage error is based on the following calibration idea of Loh (1987). Let $I_n = I_n(1-\alpha)$ be any interval with nominal level $1-\alpha$, such as one given by the bootstrap, or a simple normal approximation. Its coverage is defined to be

$$C_n(1-\alpha, F) = P_F\{\theta(F) \in I_n(1-\alpha)\} .$$

We can estimate $C_n(1-\alpha, F)$ by its bootstrap counterpart $C_n(1-\alpha, \hat{F}_n)$. Then, determine $\hat{\alpha}_n$ to satisfy

$$C_n(1-\hat{\alpha}_n, \hat{F}_n) = 1-\alpha ,$$

so that $\hat{\alpha}_n$ is the value that results in the estimated coverage to be the nominal level. The calibrated interval then is defined to be $I_n(1-\hat{\alpha}_n)$.

To fix ideas, suppose $I_n(1-\alpha)$ is the one-sided normal theory interval $(-\infty, \bar{X}_n + n^{-1/2}\hat{\sigma}_n z_{1-\alpha}]$. We argued its coverage error is $O(n^{-1/2})$. More specifically,

$$C_n(1-\alpha, F) = P_F\{n^{1/2}(\bar{X}_n - \theta(F))/\hat{\sigma}_n < z_\alpha\}$$

$$= 1 - \alpha + \frac{1}{6}\varphi(z_\alpha)(2z_\alpha^2 + 1)n^{-1/2} + O(n^{-1}) .$$

Under smoothness and moment assumptions, the bootstrap estimated coverage satisfies

$$C_n(1-\alpha, \hat{F}_n) = 1 - \alpha + \frac{1}{6}\varphi(z_\alpha)\gamma(\hat{F}_n)(2z_\alpha^2 + 1)n^{-1/2} + O_P(n^{-1}) ,$$

and the value of $\hat{\alpha}_n$ is obtained by setting the estimated coverage equal to $1-\alpha$. One can then show that

$$\hat{\alpha}_n - \alpha = -\frac{1}{6}\varphi(z_\alpha)\gamma(F)(2z_\alpha^2 + 1)n^{-1/2} + O_P(n^{-1}) . \tag{18.36}$$

By using this expansion and (18.19), it can be shown that the interval $I_n(1-\hat{\alpha}_n)$ has coverage $1-\alpha+O(n^{-1})$, and hence is second-order accurate (Problem 18.25). Thus, calibration reduces the order of coverage error.

Other Bootstrap Methods. There are now many variations on the basic bootstrap idea that yield confidence regions that are second-order accurate, assuming the validity of Edgeworth Expansions like the ones used in this section. The calibration method described above is due to Loh (1987, 1991) and is essentially equivalent to Beran's (1987, 1988b) method of prepivoting (Problem 18.29). Given an interval $I_n(1-\alpha)$ of nominal level $1-\alpha$, calibration produces a new interval, say $I_n^1(1-\alpha) = I_n(1-\hat{\alpha}_n)$, where $\hat{\alpha}_n$ is chosen by calibration. It is tempting to iterate this idea to further reduce coverage error. That is, now calibrate I_n^1 to yield a new interval I_n^2, and so on. Further reduction in coverage error is indeed possible (at the expense of increased computational effort). For further details on these and other methods such as Efron's BC_a method, see Hall and Martin (1988), Hall (1992) and Efron and Tibshirani (1993).

The analysis of this section was limited to methods for constructing confidence intervals for a mean, assuming the underlying distribution is smooth and has sufficiently many moments. But, many of the conclusions extend to smooth functions of means studied in Example 18.3.3. In particular, in order to reduce coverage error, it is desirable to use a root that is at least asymptotically pivotal, such as a studentized root that is asymptotically standard normal. Otherwise, the basic bootstrap interval (18.3) has the same order of coverage error as one based on approximating the asymptotic distribution. However, whether or not the root is asymptotically pivotal, bootstrap calibration reduces the order of coverage error. Of course, some qualifications are necessary. For one, even in the context of the mean, Cramér's condition may not hold, as in the context of a binomial proportion. Edgeworth expansions for such discrete distributions supported on a lattice are studied in Chapter 5 of Bhattacharya and Rao (1976) and Kolassa and McCullagh (1990); also see Brown et al. (2001), who study the binomial case. In other problems where smoothness is assumed, such as inference for a density or quantiles, Edgeworth expansions for appropriate statistics behave somewhat differently than they do for a mean. Such problems are treated in Hall (1992).

18.5 Hypothesis Testing

In this section, we consider the use of the bootstrap for the construction of hypothesis tests. Assume the data X^n is generated from some unknown law P. The null hypothesis H asserts that P belongs to a certain family of distributions $\mathbf{P_0}$, while the alternative hypothesis K asserts that P belongs to a family $\mathbf{P_1}$. Of course, we assume the intersection of $\mathbf{P_0}$ and $\mathbf{P_1}$ is the empty set, and the unknown law P belongs to $\mathbf{P}$, the union of $\mathbf{P_0}$ and $\mathbf{P_1}$.

There are several approaches one can take to construct a hypothesis test. First, consider the case when the null hypothesis can be expressed as a hypothesis about a real- or vector-valued parameter $\theta(P)$. Then, one can exploit the familiar duality between confidence regions and hypothesis tests to test hypotheses about $\theta(P)$. Thus, a consistent in level test of the null hypothesis that $\theta(P) = \theta_0$ can be constructed by a consistent in level confidence region for $\theta(P)$ by the rule: accept the null hypothesis if and only if the confidence region includes θ_0. Therefore, all the methods we have thus far discussed for constructing confidence regions may be utilized: methods based on a pivot, an asymptotic pivot, an asymptotic approximation, or the bootstrap. Indeed, this was the bootstrap approach already considered in Corollary 18.3.1, and it is also the basis for the multiple test construction in Section 18.6.

Example 18.5.1 (**Testing Moment Inequalities Using Bootstrap**) Consider a nonparametric version of the moment inequality testing problem in Example 14.4.8. Assume the random vectors $X_1, \ldots, X_n$ are i.i.d. P in $\mathbb{R}$ with unknown invertible covariance matrix $\Sigma = \Sigma(P)$ and mean vector $\theta = \theta(P)$. The problem is to test the null hypothesis

$$H_0 : \theta_i(P) \le 0 \quad \text{for all } i = 1, \ldots, k .$$

One can exploit the duality between confidence sets and testing to first construct a bootstrap joint "lower" confidence set for $\theta(P)$ as follows. Let $\hat{P}_n$ be the empirical measure, let $\bar{X}$ be the sample mean vector, let $\hat{\Sigma} = \Sigma(\hat{P}_n)$ be the sample covariance matrix, and let $\hat{\sigma}^2_{n,i}$ be the ith diagonal entry of $\hat{\Sigma}$. Consider the root

$$J_n(x, P) = P\left\{ \max_{1 \le i \le k} \frac{\bar{X}_{n,i} - \theta_i(P)}{\hat{\sigma}_{n,i}} \le x \right\} .$$

This leads to the bootstrap confidence set

$$\left\{\theta : \max_{1\le i\le k} \frac{\bar{X}_{n,i}-\theta_i}{\hat{\sigma}_{n,i}} \le J_n^{-1}(1-\alpha, \hat{P}_n)\right\} ,$$

which yields the joint lower confidence set as

$$[\bar{X}_{n,i} - \hat{\sigma}_{n,i} J_n^{-1}(1-\alpha, \hat{P}_n), \infty) .$$

These k semi-infinite intervals simultaneously contain the true $\theta_i(P)$ with asymptotic probability $1-\alpha$. The argument is very similar to that in Example 18.3.4.

Returning to the moment inequality testing problem, a solution is to reject the null hypothesis H_0 if the 0 vector is not included in the joint confidence set. Such a method asymptotically controls the probability of a Type 1 error. The power of this test is considered in Problem 18.30. ■

However, not all hypothesis testing problems fit nicely into the framework of testing parameters. For example, consider the problem of testing whether the data come from a certain parametric submodel (such as the family of normal distributions) of a nonparametric model, the so-called goodness of fit problem. Or, when X_i is vector-valued, consider the problem of testing whether X_i has a distribution that is spherically symmetric.

Given a test statistic T_n, its distribution must be known, estimated, or approximated (at least under the null hypothesis), in order to construct a critical value. The approach taken in this section is to estimate the null distribution of T_n by resampling from a distribution obeying the constraints of the null hypothesis.

To be explicit, assume we wish to construct a test based on a real-valued test statistic $T_n = T_n(X^n)$ which is consistent in level and power. Large values of T_n reject the null hypothesis. Thus, having picked a suitable test statistic T_n, our goal is to construct a critical value, say $c_n(1-\alpha)$, so that the test which rejects if and only if T_n exceeds $c_n(1-\alpha)$ satisfies

$$P\{T_n(X^n) > c_n(1-\alpha)\} \to \alpha \text{ as } n \to \infty$$

when $P \in \mathbf{P_0}$. Furthermore, we require this rejection probability to tend to one when $P \in \mathbf{P_1}$. Unlike the classical case, the critical value will be constructed to be data-dependent (as in the case of a permutation test). To see how the bootstrap can be used to determine a critical value, let the distribution of T_n under P be denoted by

$$G_n(x, P) = P\{T_n(X^n) \le x\} .$$

Note that we have introduced $G_n(\cdot, P)$ instead of utilizing $J_n(\cdot, P)$ to distinguish from the case of confidence intervals where $J_n(\cdot, P)$ represents the distribution of a root which may depend both on the data and on P. In the hypothesis testing context, $G_n(\cdot, P)$ represents the distribution of a statistic (and not a root) under P. Let

$$g_n(1-\alpha, P) = \inf\{x : G_n(x, P) \geq 1-\alpha\} .$$

Typically, $G_n(\cdot, P)$ will converge in distribution to a limit law $G(\cdot, P)$, whose $1-\alpha$ quantile is denoted by $g(1-\alpha, P)$.

The bootstrap approach is to estimate the null sampling distribution by $G_n(\cdot, \hat{Q}_n)$, where $\hat{Q}_n$ is an estimate of P in $\mathbf{P_0}$ so that $\hat{Q}_n$ satisfies the constraints of the null hypothesis, since critical values should be determined as if the null hypothesis were true. A bootstrap critical value can then be defined by $g_n(1-\alpha, \hat{Q}_n)$. The resulting nominal level α bootstrap test rejects H if and only if $T_n > g_n(1-\alpha, \hat{Q}_n)$.

Notice that we would not want to replace a $\hat{Q}_n$ satisfying the null hypothesis constraints by the empirical distribution function $\hat{P}_n$, the usual resampling mechanism of resampling the data with replacement. One might say that the bootstrap is so adept at estimating the distribution of a statistic that $G_n(\cdot, \hat{P}_n)$ is a good estimate of $G_n(\cdot, P)$ whether or not P satisfies the null hypothesis constraints. Hence, the test that rejects when T_n exceeds $g_n(1-\alpha, \hat{P}_n)$ will (under suitable conditions) behave asymptotically like the test that rejects when T_n exceeds $g_n(1-\alpha, P)$, and this test has an asymptotic probability of α of rejecting the null hypothesis, even if $P \in \mathbf{P_1}$. But, when $P \in \mathbf{P_1}$, we would want the test to reject with probability that is approaching one.

Thus, the choice of resampling distribution $\hat{Q}_n$ should satisfy the following. If $P \in \mathbf{P_0}$, $\hat{Q}_n$ should be near P so that $G_n(\cdot, P) \approx G_n(\cdot, \hat{Q}_n)$; then, $g_n(1-\alpha, P) \approx g_n(1-\alpha, \hat{Q}_n)$ and the asymptotic rejection probability approaches α. If, on the other hand, $P \in \mathbf{P_1}$, $\hat{Q}_n$ should not approach P, but some P_0 in $\mathbf{P_0}$. In this way, the critical value should satisfy

$$g_n(1-\alpha, \hat{Q}_n) \approx g_n(1-\alpha, P_0) \to g(1-\alpha, P_0) < \infty$$

as $n \to \infty$. Then, assuming the test statistic is constructed so that $T_n \to \infty$ under P when $P \in \mathbf{P_1}$, we will have

$$P\{T_n > g_n(1-\alpha, \hat{Q}_n)\} \approx P\{T_n > g(1-\alpha, P_0)\} \to 1$$

as $n \to \infty$, by Slutsky's Theorem.

As in the construction of confidence intervals, $G_n(\cdot, P)$ must be smooth in P in order for the bootstrap to succeed. In the theorem below, rather than specifying a set of sequences $\mathbf{C}_P$ as was done in Theorem 18.3.1, smoothness is described in terms of a metric d, but either approach could be used. The proof is analogous to the proof of Theorem 18.3.1.

Theorem 18.5.1 *Let X^n be generated from a probability law $P \in \mathbf{P_0}$. Assume the following triangular array convergence: $d(P_n, P) \to 0$ and $P \in \mathbf{P_0}$ implies $G_n(\cdot, P_n)$ converges weakly to $G(\cdot, P)$ with $G(\cdot, P)$ continuous. Moreover, assume $\hat{Q}_n$ is an estimator of P based on X^n which satisfies $d(\hat{Q}_n, P) \to 0$ in probability whenever $P \in \mathbf{P_0}$. Then,*

$$P\{T_n > g_n(1-\alpha, \hat{Q}_n)\} \to \alpha \quad \text{as } n \to \infty .$$

Example 18.5.2 (**Normal Correlation**) Suppose (Y_i, Z_i), $i = 1, \ldots, n$ are i.i.d. bivariate normal with unknown means, variances, and correlation ρ. The null hypothesis specifies $\rho = \rho_0$ versus $\rho > \rho_0$. Let $T_n = n^{1/2}\hat{\rho}_n$, where $\hat{\rho}_n$ is the usual sample correlation. Under the null hypothesis, the distribution of T_n doesn't depend on any unknown parameters. So, if $\hat{Q}_n$ is any bivariate normal distribution with $\rho = \rho_0$, the bootstrap sampling distribution $G_n(\cdot, \hat{Q}_n)$ is exactly equal to the true null sampling distribution. Note, however, that inverting a parametric bootstrap confidence bound using the root $n^{1/2}(\hat{\rho}_n - \rho)$ would not be exact. ■

Example 18.5.3 (**Likelihood Ratio Tests**) Suppose $X_1, \ldots, X_n$ are i.i.d. according to a model $\{P_\theta,\ \theta \in \Omega\}$, where Ω is an open subset of $\mathbb{R}^k$. Assume θ is partitioned as (ξ, μ), where ξ is a vector of length p and μ is a vector of length $k - p$. The null hypothesis parameter space Ω_0 specifies $\xi = \xi_0$. Under the conditions of Theorem 14.4.2, the likelihood ratio statistic $T_n = 2\log(R_n)$ is asymptotically χ^2_p under the null hypothesis. Suppose $(\xi_0, \hat{\mu}_{n,0})$ is an efficient likelihood estimator of θ for the model Ω_0. Rather than using the critical value obtained from χ^2_p, one could bootstrap T_n. So, let $G_n(x, \theta)$ denote the distribution of T_n under θ. An appropriate parametric bootstrap test obeying the null hypothesis constraints is to reject the null when T_n exceeds the $1 - \alpha$ quantile of $G_n(x, (\xi_0, \hat{\mu}_{n,0}))$. Beran and Ducharme (1991) argue that, under regularity conditions, the bootstrap test has error in rejection probability equal to $O(n^{-2})$, while the usual likelihood ratio test has error $O(n^{-1})$. Moreover, the bootstrap test can be viewed as an analytical approximation to a Bartlett-corrected likelihood ratio test (see Section 14.4.4). In essence, the bootstrap automatically captures the Bartlett correction and avoids the need for analytical calculation. As an example, recall Example 14.4.7, where it was observed the Bartlett-corrected likelihood ratio test has error $O(n^{-2})$. Here, the bootstrap test is exact (Problem 18.33). ■

Example 18.5.4 (**Behrens–Fisher Problem Revisited**) For $j = 1, 2$, let $X_{i,j}$, $i = 1, \ldots, n_j$ be independent with $X_{i,j}$ distributed as $N(\mu_j, \sigma_j^2)$. All four parameters are unknown and vary independently. The null hypothesis asserts $\mu_1 = \mu_2$ and the alternative is $\mu_1 > \mu_2$. Let $n = n_1 + n_2$, and for simplicity assume n_1 to be the integer part of λn for some $0 < \lambda < 1$. Let $(\bar{X}_{n,j}, S^2_{n,j})$ be the usual unbiased estimators of (μ_j, σ_j^2) based on the jth sample. Consider the test statistic

$$T_n = (\bar{X}_1 - \bar{X}_2)/\sqrt{\frac{S^2_{n,1}}{n_1} + \frac{S^2_{n,2}}{n_2}}\ .$$

By Example 15.5.4, the test that rejects the null hypothesis when $T_n > z_{1-\alpha}$ is efficient. However, we now study its actual rejection probability.

The null distribution of T_n depends only on $\sigma^2 = (\sigma_1^2, \sigma_2^2)$ through the ratio σ_1/σ_2, and we denote this distribution by $G_n(\cdot, \sigma^2)$. Let $S_n^2 = (S^2_{n,1}, S^2_{n,2})$. Like the method used in Problem 13.28, by conditioning on S_n^2, we can write

$$G_n(x, \sigma^2) = E[a(S_n^2, \sigma^2, x)]\ ,$$

where

$$a(S_n^2, \sigma^2, x) = \Phi[(1+\delta)^{1/2} x]$$

and

$$\delta = \sum_{j=1}^{2} n_j^{-1}(S_{n,j}^2 - \sigma_j^2) / \sum_{j=1}^{2} n_j^{-1}\sigma_j^2 \ .$$

By Taylor expansion and the moments of S_n^2, it follows that (Problem 18.34)

$$G_n(x, \sigma^2) = \Phi(x) + \frac{1}{n} b_n(x, \sigma^2) + O(n^{-2}) \ , \tag{18.37}$$

where

$$\frac{1}{n} b_n(x, \sigma^2) = -(x + x^3)\phi(x)\rho_n^2/4$$

is $O(n^{-1})$ and

$$\rho_n^2 = \sum_{j=1}^{2} (n_j - 1)^{-1} n_j^{-2} \sigma_j^4 / (\sum_{j=1}^{2} n_j^{-1}\sigma_j^2)^2 \ .$$

Correspondingly, the quantile function satisfies

$$G_n^{-1}(1-\alpha, \sigma^2) = z_{1-\alpha} + (z_{1-\alpha} + z_{1-\alpha}^3)\rho_n^2/4 + O(n^{-2}) \ . \tag{18.38}$$

It follows that the rejection probability of the asymptotic test that rejects when $T_n > z_{1-\alpha}$ is $\alpha + O(n^{-1})$.

Consider next the (parametric) bootstrap-t, which rejects when $T_n > G_n^{-1}(1-\alpha, S_n^2)$. Its rejection probability can be expressed as

$$1 - E[a(S_n^2, \sigma^2, G_n^{-1}(1-\alpha, S_n^2))] \ .$$

By Taylor expansion, it can be shown that the rejection probability of the test is $\alpha + O(n^{-2})$ (Problem 18.35). Thus, the bootstrap-t improves upon the asymptotic expansion. In fact, bootstrap calibration (or the use of prepivoting) further reduces the error in rejection probability to $O(n^{-3})$. Details are in Beran (1988), who further argues that the Welch method described in Section 13.2.1 behaves like the bootstrap-t method. Although the Welch approximation is based on elegant mathematics, the bootstrap approach essentially reproduces the analytical approximation automatically. ■

Example 18.5.5 (**Nonparametric Mean**) Let $X_1, \ldots, X_n$ be i.i.d. observations on the real line with probability law P, mean $\mu(P)$ and finite variance $\sigma^2(P)$. The problem to test $\mu(P) = 0$ against either a one-sided or two-sided alternative. Let $\mathbf{P_0}$ be the set of distributions with mean zero and finite variance. In the one-sided case, consider the test statistic $T_n = n^{1/2}\bar{X}_n$, where $\bar{X}_n$ is the sample mean, since test statistics based on $\bar{X}_n$ were seen in Section 13.4 to possess a certain optimality property. We will also consider the studentized statistic $T_n' = n^{1/2}\bar{X}_n/S_n$, where we shall take S_n^2 to be the unbiased estimate of variance. To apply Theorem 18.5.1, let $\hat{Q}_n$ be the empirical distribution $\hat{P}_n$ shifted by $\bar{X}_n$ so it has mean 0. Then, the error in rejection probability will be $O(n^{-1/2})$ for T_n, and will be $O(n^{-1})$ for T_n', at least under the assumptions that F is smooth and has infinitely many moments; these statements follow from the results in Section 18.4 (Problem 18.37).

While shifting the empirical distribution works in this example, it is not easy to generalize when testing other parameters. Therefore, we consider the following alternative approach. The idea is to choose the distribution in $\mathbf{P_0}$ that is in some sense closest to the empirical distribution $\hat{P}_n$. One way to describe closeness is the following. For distributions P and Q on the real line, let $\delta_{KL}(P, Q)$ be the (forward) Kullback–Leibler divergence between P and Q (studied in Example 11.3.1), defined by

$$\delta_{KL}(P, Q) = \int log(\frac{dP}{dQ})dP \ . \tag{18.39}$$

Note that $\delta_{KL}(P, Q)$ may be ∞, δ_{KL} is not a metric, and it is not even symmetric in its arguments. Let $\hat{Q}_n$ be the Q that minimizes $\delta_{KL}(\hat{P}_n, Q)$ over Q in $\mathbf{P_0}$. This choice for $\hat{Q}_n$ can be shown to be well-defined and corresponds to finding the nonparametric maximum likelihood estimator of P assuming P is constrained to have mean zero. (Another possibility is to minimize the (backward) Kullback–Leibler divergence $\delta_{KL}(Q, \hat{P}_n)$.) By Efron (1981) (Problem 18.38), $\hat{Q}_n$ assigns mass w_i to X_i, where w_i satisfies

$$w_i \propto \frac{(1+tX_i)^{-1}}{\sum_{j=1}^n (1+tX_j)^{-1}}$$

and t is chosen so that $\sum_{i=1}^n w_i X_i = 0$. Now, one could bootstrap either T_n or T_n' from $\hat{Q}_n$.

In fact, this approach suggests an alternative test statistic given by $T_n'' = n\delta_{KL}(\hat{P}_n, \hat{Q}_n)$, where $\hat{Q}_n$ is the Q minimizing the Kullback–Leibler divergence $\delta_{KL}(\hat{P}_n, Q)$ over Q in $\mathbf{P_0}$. This is equivalent to the test statistic used by Owen (1988, 2001) in his construction of empirical likelihood, who shows the limiting distribution of $2T_n''$ under the null hypothesis is Chi-squared with 1 degree of freedom. The wide scope of empirical likelihood is presented in Owen (2001). ■

Example 18.5.6 (**Goodness of fit**) The problem is to test whether the underlying probability distribution P belongs to a parametric family of distributions $\mathbf{P_0} = \{P_\theta, \theta \in \Theta_0\}$, where Θ_0 is an open subset of k-dimensional Euclidean space. Let $\hat{P}_n$ be the empirical measure based on $X_1, \ldots, X_n$. Let $\hat{\theta}_n \in \Theta_0$ be an estimator of θ. Consider the test statistic

$$T_n = n^{1/2}\delta(\hat{P}_n, P_{\hat{\theta}_n}) ,$$

where δ is some measure (typically a metric) between $\hat{P}_n$ and $P_{\hat{\theta}_n}$. (In fact, δ need not even be symmetric, which is useful sometimes: for example, consider the Cramér–von Mises statistic.) Beran (1986) considers the case where $\hat{\theta}_n$ is a minimum distance estimator, while Romano (1988) assumes that $\hat{\theta}_n$ is some asymptotically linear estimator (like an efficient likelihood estimator). For the resampling mechanism, take $\hat{Q}_n = P_{\hat{\theta}_n}$. Beran (1986) and Romano (1988) give different sets of conditions so that the above theorem is applicable, both requiring the machinery of empirical processes. ■

Example 18.5.7 (**Moment inequalties**) Consider testing moment inequalities as in Example 18.5.1 based on the test statistic

$$T_n = \max_{1 \le i \le k} \frac{\bar{X}_{n,i}}{\hat{\sigma}_{n,i}} .$$

To simplify the point of the example, assume a parametric model with $P = P_\theta$ multivariate normal with unknown mean vector θ and known covariance matrix Σ. In order to apply Theorem 18.5.1, a reasonable choice for resampling distribution under the null hypothesis would be $\hat{Q}_n = P_{\hat{\mu}_n}$, where $\hat{\mu}_n$ is an estimator of θ under H_0. If we further assume Σ is the identity, then a reasonable choice for $\hat{\mu}_n$ is the maximum likelihood estimator under the null hypothesis constraint; so, $\hat{\mu}_n$ has ith component $\hat{\mu}_{n,i} = \min(\bar{X}_{n,i}, 0)$. Then, the conditions in Theorem 18.5.1 do not hold, and the bootstrap is too liberal; see Problem 18.31. Intuitively, the level of the test would be controlled by using a critical value based on the distribution of the test statistic when $\theta = 0$. Instead, the bootstrap procedure sometimes uses a critical value based on the distribution of the test statistic under $\hat{\mu}_n$, which is component-wise no bigger than $\bar{X}_n$ and therefore leads to a smaller critical value. ■

18.6 Stepdown Multiple Testing

Suppose data $X = X^n$ is generated from some unknown probability distribution P, where P belongs to a certain family of probability distributions Ω. For $j = 1, \ldots, s$, consider the problem of simultaneously testing hypotheses $H_j : P \in \omega_j$.

For any subset $K \subseteq \{1, \ldots, s\}$, let $H_K = \bigcap_{j \in K} H_j$ be the hypothesis that $P \in \bigcap_{j \in K} \omega_j$. Suppose that a test of the individual hypothesis H_j is based on a test statistic $T_{n,j}$, with large values indicating evidence against the H_j.

The goal is to construct a stepdown method that controls the familywise error rate (FWER). Recall that the FWER is the probability of rejecting at least one true null hypothesis. More specifically, if P is the true probability mechanism, let $I = I(P) \subseteq \{1, \ldots, s\}$ denote the indices of the set of true hypotheses; that is, $i \in I$ if and only $P \in \omega_i$. Then, FWER is the probability under P that any H_i with $i \in I$ is rejected. To show its dependence on P, we may write FWER = FWER$_P$. We require that any procedure satisfy that the FWER be no bigger than α (at least asymptotically).

Suppose H_i is specified by a real-valued parameter $\beta_i(P) = 0$. Then, one approach to constructing a multiple test is to invert a simultaneous confidence region. Under the setup of Example 18.3.4, with $\beta_i(P) = f_i(\theta(P))$, any hypothesis H_i is rejected if $f_i(\hat{\theta}_n) > \hat{b}_n(1-\alpha)$. A procedure that uses a common critical value $\hat{b}_n(1-\alpha)$ for all the hypotheses is called a single-step method.

Another approach is to compute (or approximate) a p-value for each individual test, and then use Holm's method discussed in Section 9.1, However, Holm's method, which makes no assumptions about the dependence structure of the test statistics, can be improved by methods that implicitly or explicitly estimate this dependence structure. In this section, we consider a stepdown procedure that incorporates the dependence structure and thereby improves upon the two methods just described.

Let

$$T_{n,r_1} \ge T_{n,r_2} \ge \cdots \ge T_{n,r_s} \tag{18.40}$$

denote the observed ordered test statistics, and let $H_{r_1}, H_{r_2}, \ldots, H_{r_s}$ be the corresponding hypotheses.

Recall the stepdown method presented in Procedure 9.1.1. The problem now is how to construct the $\hat{c}_{n,K}(1-\alpha)$ so that the FWER is controlled, at least asymptotically. The following is an immediate consequence of Theorem 9.1.3, and reduces the multiple testing problem of asymptotically controlling the FWER to the single testing problem of asymptotically controlling the probability of a Type 1 error.

Corollary 18.6.1 *Let P denote the true distribution generating the data. Consider Procedure 9.1.1 based on critical values $\hat{c}_{n,K}(1-\alpha)$ which satisfy the monotonicity requirement: for any $K \supseteq I(P)$,*

$$\hat{c}_{n,K}(1-\alpha) \ge \hat{c}_{n,I(P)}(1-\alpha)\ . \tag{18.41}$$

If $\hat{c}_{n,I(P)}(1-\alpha)$ satisfies

$$\limsup_n P\{\max(T_{n,j}\ :\ j \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha)\} \le \alpha\ , \tag{18.42}$$

then $\limsup_n FWER_P \le \alpha$ *as $n \to \infty$.*

Under the monotonicity requirement (18.41), the multiplicity problem is effectively reduced to testing a single intersection hypothesis at a time. So, the problem now is to construct intersection tests whose critical values are monotone and asymptotically control the rejection probability.

We now specialize a bit and develop a concrete construction based on the bootstrap. Suppose hypothesis H_i is specified by $\{P\ :\ \theta_i(P) = 0\}$ for some real-valued parameter θ_i, and $\hat{\theta}_{n,i}$ is an estimate of θ_i. Also, let $T_{n,i} = \tau_n|\hat{\theta}_{n,i}|$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$; usually, $\tau_n = n^{1/2}$. The bootstrap method relies on its ability to approximate the joint distribution of $\{\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]\ :\ i \in K\}$, whose distribution we denote by $J_{n,K}(P)$. Also, let $L_{n,K}(P)$ denote the distribution under P of $\max\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)|\ :\ i \in K\}$, with corresponding distribution function

$L_{n,K}(x, P)$ and α-quantile

$$b_{n,K}(\alpha, P) = \inf\{x : L_{n,K}(x, P) \geq \alpha\} .$$

Let $\hat{Q}_n$ be some estimate of P. Then, a nominal $1 - \alpha$ level bootstrap confidence region for the subset of parameters $\{\theta_i(P) : i \in K\}$ is given by

$$\{(\theta_i : i \in K) : \max_{i \in K} \tau_n |\hat{\theta}_{n,i} - \theta_i| \leq b_{n,K}(1 - \alpha, \hat{Q}_n)\} .$$

So a value of 0 for $\theta_i(P)$ falls outside the region iff

$$T_{n,i} = \tau_n |\hat{\theta}_{n,i}| > b_{n,K}(1 - \alpha, \hat{Q}_n) .$$

By the usual duality of confidence sets and hypothesis tests, this suggests the use of the critical value

$$\hat{c}_{n,K}(1 - \alpha) = b_{n,K}(1 - \alpha, \hat{Q}_n) , \tag{18.43}$$

at least if the bootstrap is a valid asymptotic approach for confidence region construction.

Note that, regardless of asymptotic behavior, the monotonicity assumption (18.41) is always satisfied for the choice (18.43). Indeed, for any Q and if $I \subseteq K$, $b_{n,I}(1 - \alpha, Q)$ is the $1 - \alpha$ quantile under Q of the maximum of $|I|$ variables, while $b_{n,K}(1 - \alpha, Q)$ is the $1 - \alpha$ quantile of these same $|I|$ variables together with $|K| - |I|$ variables.

Therefore, in order to apply Theorem 18.6.1 to conclude $\limsup_n \text{FWER}_P \leq \alpha$, it is now only necessary to study the asymptotic behavior of $b_{n,K}(1 - \alpha, \hat{Q}_n)$ in the case $K = I(P)$. For this, we assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that $\theta_i(P) = 0$ for all $i \in I(P)$; that is, we assume the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ for $i \in I(P)$. In particular, we assume

$$J_{n,I(P)}(P) \xrightarrow{d} J_{I(P)}(P) , \tag{18.44}$$

a nondegenerate limit law. Assumption (18.44) implies $L_{n,I(P)}(P)$ has a limiting distribution $L_{I(P)}(P)$, with c.d.f. denoted by $L_{I(P)}(x, P)$. We will further assume $L_{I(P)}(P)$ is continuous and strictly increasing on its support. It follows that

$$b_{n,I(P)}(1 - \alpha, P) \to b_{I(P)}(1 - \alpha, P) , \tag{18.45}$$

where $b_{I(P)}(\alpha, P)$ is the α-quantile of the limiting distribution $L_{I(P)}(P)$.

Theorem 18.6.1 *Fix P and assume (18.44) and that $L_{I(P)}(P)$ is continuous and strictly increasing on its support. Let $\hat{Q}_n$ be an estimate of P satisfying: for any metric ρ metrizing weak convergence on $\mathbb{R}^{|I(P)|}$,*

$$\rho\left(J_{n,I(P)}(P),\, J_{n,I(P)}(\hat{Q}_n)\right) \xrightarrow{P} 0 \; . \qquad (18.46)$$

Consider the generic stepdown method in Procedure 9.1.1 with $c_{n,K}(1-\alpha)$ *equal to* $b_{n,K}(1-\alpha, \hat{Q}_n)$*. Then,* $\limsup_n FWER_P \le \alpha$*.*

Proof. By the Continuous Mapping Theorem and a subsequence argument (Problem 18.40), the assumption (18.44) implies

$$\rho_1\left(L_{n,I(P)}(P),\, L_{n,I(P)}(\hat{Q}_n)\right) \xrightarrow{P} 0 \; , \qquad (18.47)$$

where ρ_1 is any metric metrizing weak convergence on $\mathbb{R}$. It follows from Problem 11.30, which is a generalization of Lemma 11.2.1, that

$$b_{n,I(P)}(1-\alpha, \hat{Q}_n) \xrightarrow{P} b_{I(P)}(1-\alpha, P) \; .$$

By Slutsky's Theorem,

$$P\{\max(T_{n,j} : \; j \in I(P))\} > b_{n,I(P)}(1-\alpha, \hat{Q}_n)\} \to 1 - L_{I(P)}(b_{I(P)}(1-\alpha, P), P),$$

and the last expression is α. ■

Example 18.6.1 (**Multivariate Mean**) Assume $X_i = (X_{i,1}, \ldots, X_{i,s})$ are n i.i.d. random vectors with $E(|X_i|^2) < \infty$ and mean vector $\mu = (\mu_1, \ldots, \mu_s)$. Note that the vector X_i can have an arbitrary s-variate distribution, so that multivariate normality is not assumed as it was in Example 9.1.7. Suppose H_i specifies $\mu_i = 0$ and $T_{n,i} = n^{-1/2}|\sum_{j=1}^n X_{j,i}|$. Then, the conditions of Theorem 18.6.1 are satisfied by Example 18.3.2. Alternatively, one can also consider the studentized test statistic $t_{n,i} = T_{n,i}/S_{n,i}$, where $S^2_{n,i}$ is the sample variance of the ith components of the data (Problem 18.41). ■

Example 18.6.2 (**Comparing Treatment Means**) For $i = 1, \ldots, k$, suppose we observe k independent samples, and the ith sample consists of n_i i.i.d. observations $X_{i,1}, \ldots, X_{i,n_i}$ with mean μ_i and finite variance σ_i^2. Hypothesis $H_{i,j}$ specifies $\mu_i = \mu_j$, so that the problem is to compare all $s = \binom{k}{2}$ means. (Note that we are indexing hypotheses and test statistics now by 2 indices i and j.) Let $T_{n,i,j} = n^{1/2}|\bar{X}_{n,i} - \bar{X}_{n,j}|$, where $\bar{X}_{n,i} = \sum_{j=1}^n X_{i,j}/n_i$. Let $\hat{Q}_{n_i,i}$ be the empirical distribution of the ith sample. The bootstrap resampling scheme is to independently resample n_i observations from $\hat{Q}_{n,i}$, $i = 1, \ldots, k$. Then, Theorem 18.6.1 applies and it also applies to appropriately studentized statistics (Problem 18.42). The setup can easily accommodate comparisons of k treatments with a control group (Problem 18.43). ■

Example 18.6.3 (**Testing Correlations**) Suppose $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^k$, so that $X_i = (X_{i,1}, \ldots, X_{i,k})$. Assume $E|X_{i,j}|^2 < \infty$ and $Var(X_{i,j}) > 0$, so that the correlation between $X_{1,i}$ and $X_{1,j}$, namely $\rho_{i,j}$ is well defined. Let $H_{i,j}$

denote the hypothesis that $\rho_{i,j} = 0$, so that the multiple testing problem consists in testing all $s = \binom{k}{2}$ pairwise correlations. Also let $T_{n,i,j}$ denote the ordinary sample correlation between variables i and j. (Note that we are indexing hypotheses and test statistics now by 2 indices i and j.) By Example 18.3.3, the conditions for the bootstrap hold because correlations are smooth functions of means. ■

18.7 Subsampling

In this section, a general theory for the construction of approximate confidence sets or hypothesis tests is presented, so the goal is the same as that of the bootstrap. The basic idea is to approximate the sampling distribution of a statistic based on the values of the statistic computed over smaller subsets of the data. For example, in the case where the data are n observations which are independent and identically distributed, a statistic $\hat{\theta}_n$ is computed based on the entire data set and is recomputed over all $\binom{n}{b}$ data sets of size b. Implicit is the notion of a statistic sequence, so that the statistic is defined for samples of size n and b. These recomputed values of the statistic are suitably normalized to approximate the true sampling distribution.

This approach based on subsamples is perhaps the most general one for approximating a sampling distribution, in the sense that consistency holds under extremely weak conditions. That is, it will be seen that, under very weak assumptions on b, the method is consistent whenever the original statistic, suitably normalized, has a limit distribution under the true model. The bootstrap, on the other hand, requires that the distribution of the statistic is somehow locally smooth as a function of the unknown model. In contrast, no such assumption is required in the theory for subsampling. Indeed, the method here is applicable even in the several known situations which represent counterexamples to the bootstrap. However, when both subsampling and the bootstrap are consistent, the bootstrap is typically more accurate.

To appreciate why subsampling behaves well under such weak assumptions, note that each subset of size b (taken without replacement from the original data) is indeed a sample of size b from the true model. If b is small compared to n (meaning $b/n \to 0$), then there are many (namely $\binom{n}{b}$) subsamples of size b available. Hence, it should be intuitively clear that one can at least approximate the sampling distribution of the (normalized) statistic $\hat{\theta}_b$ by recomputing the values of the statistic over all these subsamples. But, under the weak convergence hypothesis, the sampling distributions based on samples of size b and n should be close. The bootstrap, on the other hand, is based on recomputing a statistic over a sample of size n from some estimated model which is hopefully close to the true model.

The use of subsample values to approximate the variance of a statistic is well known. The Quenouille-Tukey jackknife estimates of bias and variance based on computing a statistic over all subsamples of size $n - 1$ has been well studied and is closely related to the mean and variance of our estimated sampling distribution with $b = n - 1$. For further history of subsampling methods, see Politis et al. (1999).

18.7.1 The Basic Theorem in the I.I.D. Case

Suppose $X_1, \dots, X_n$ is a sample of n i.i.d. random variables taking values in an arbitrary sample space S. The common probability measure generating the observations is denoted by P. The goal is to construct a confidence region for some parameter $\theta(P)$. For now, assume θ is real-valued, but this can and will be generalized to allow for the construction of confidence regions for multivariate parameters or confidence bands for functions.

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ be an estimator of $\theta(P)$. It is desired to estimate the true sampling distribution of $\hat{\theta}_n$ in order to make inferences about $\theta(P)$. Nothing is assumed about the form of the estimator.

As in previous sections, let $J_n(P)$ be the sampling distribution of the root $\tau_n(\hat{\theta}_n - \theta(P))$ based on a sample of size n from P, where τ_n is a normalizing constant. Here, τ_n is assumed known and does not depend on P. Also define the corresponding cumulative distribution function:

$$J_n(x, P) = P\{\tau_n[\hat{\theta}_n(X_1, \dots, X_n) - \theta(P)] \le x\} .$$

Essentially, the only assumption that we will need to construct asymptotically valid confidence intervals for $\theta(P)$ is the following.

Assumption 18.7.1 There exists a limiting distribution $J(P)$ such that $J_n(P)$ converges weakly to $J(P)$ as $n \to \infty$.

This assumption will be required to hold for some sequence τ_n. The most informative case occurs when τ_n is such that the limit law $J(P)$ is nondegenerate.

To describe the subsampling method, consider the $N_n = \binom{n}{b}$ subsets of size b of the data $\{X_1, \dots, X_n\}$; call them $Y_1, \dots, Y_{N_n}$, ordered in any fashion. Thus, each Y_i constitutes a sample of size b from P. Of course, the Y_i depend on b and n, but this notation has been suppressed. Only a very weak assumption on b will be required. In the consistency results that follow, it will be assumed that $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Now, let $\hat{\theta}_{n,b,i}$ be equal to the statistic $\hat{\theta}_b$ evaluated at the data set Y_i. The approximation to $J_n(x, P)$ we study is defined by

$$L_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \le x\} . \tag{18.48}$$

The motivation behind the method is the following. For any i, Y_i is actually a random sample of b i.i.d. observations from P. Hence, the *exact* distribution of $\tau_b(\hat{\theta}_{n,b,i} - \theta(P))$ is $J_b(P)$. The empirical distribution of the N_n values of $\tau_b(\hat{\theta}_{n,b,i} - \theta(P))$ should then serve as a good approximation to $J_n(P)$. Of course, $\theta(P)$ is unknown, so we replace $\theta(P)$ by $\hat{\theta}_n$, which is asymptotically permissible because $\tau_b(\hat{\theta}_n - \theta(P))$ is of order τ_b/τ_n, and τ_b/τ_n will be assumed to tend to zero.

Theorem 18.7.1 *Suppose Assumption 18.7.1 holds. Also, assume $\tau_b/\tau_n \to 0$, $b \to \infty$, and $b/n \to 0$ as $n \to \infty$.*

(i) If x is a continuity point of $J(\cdot, P)$, then $L_{n,b}(x) \to J(x, P)$ in probability.

(ii) If $J(\cdot, P)$ is continuous, then

$$\sup_x |L_{n,b}(x) - J_n(x, P)| \to 0 \textit{ in probability} . \tag{18.49}$$

(iii) Let

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geq 1-\alpha\} .$$

and

$$c(1-\alpha, P) = \inf\{x : J(x, P) \geq 1-\alpha\} .$$

If $J(\cdot, P)$ is continuous at $c(1-\alpha, P)$, then

$$P\{\tau_n[\hat{\theta}_n - \theta(P)] \leq c_{n,b}(1-\alpha)\} \to 1-\alpha \textit{ as } n \to \infty . \tag{18.50}$$

Therefore, the asymptotic coverage probability under P of the confidence interval $[\hat{\theta}_n - \tau_n^{-1} c_{n,b}(1-\alpha), \infty)$ is the nominal level $1-\alpha$.

Proof. Let

$$U_n(x) = U_{n,b}(x, P) = N_n^{-1} \sum_{i=1}^{N_n} I\{\tau_b[\hat{\theta}_{n,b,i} - \theta(P)] \leq x\} . \tag{18.51}$$

Note that the dependence of $U_n(x)$ on b and P will now be suppressed for notational convenience. To prove (i), it suffices to show $U_n(x)$ converges in probability to $J(x, P)$ for every continuity point x of $J(x, P)$. To see why, note that

$$L_{n,b}(x) = N_n^{-1} \sum_i I\{\tau_b[\hat{\theta}_{n,b,i} - \theta(P)] + \tau_b[\theta(P) - \hat{\theta}_n] \leq x\} ,$$

so that for every $\epsilon > 0$,

$$U_n(x-\epsilon) I\{E_n\} \leq L_{n,b}(x) I\{E_n\} \leq U_n(x+\epsilon) I\{E_n\} ,$$

where $I\{E_n\}$ is the indicator of the event $E_n \equiv \{\tau_b |\theta(P) - \hat{\theta}_n| \leq \epsilon\}$. But, the event E_n has probability tending to one. So, with probability tending to one,

$$U_n(x-\epsilon) \leq L_{n,b}(x) \leq U_n(x+\epsilon)$$

for any $\epsilon > 0$. Hence, if $x+\epsilon$ and $x-\epsilon$ are continuity points of $J(\cdot, P)$, then $U_n(x \pm \epsilon) \to J(x \pm \epsilon, P)$ in probability implies

$$J(x-\epsilon, P)-\epsilon \le L_{n,b}(x) \le J(x+\epsilon, P)+\epsilon$$

with probability tending to one. Now, let $\epsilon \to 0$ so that $x \pm \epsilon$ are continuity points of $J(\cdot, P)$. Then, it suffices to show $U_n(x) \to J(x, P)$ in probability for all continuity points x of $J(\cdot, P)$. But, $0 \le U_n(x) \le 1$ and

$$E[U_n(x)] = J_b(x, P) .$$

Since $J_b(x, P) \to J(x, P)$, it suffices to show $Var[U_n(x)] \to 0$. To this end, suppose k is the greatest integer less than or equal to n/b. For $j = 1, \ldots, k$, let $R_{n,b,j}$ be equal to the statistic $\hat{\theta}_b$ evaluated at the data set $\hat{\theta}_b(X_{b(j-1)+1}, X_{b(j-1)+2}, \ldots, X_{b(j-1)+b})$ and set

$$\bar{U}_n(x) = k^{-1}\sum_{j=1}^{k} I\{\tau_b[R_{n,b,j}-\theta(P)] \le x\} .$$

Clearly, $\bar{U}_n(x)$ and $U_n(x)$ have the same expectation. But, since $\bar{U}_n(x)$ is the average of k i.i.d. variables (each of which is bounded between 0 and 1), it follows that

$$Var[\bar{U}_n(x)] \le \frac{1}{4k} \to 0$$

as $n \to \infty$. Intuitively, $U_n(x)$ should have a smaller variance than $\bar{U}_n(x)$, because $\bar{U}_n(x)$ uses the ordering in the sample in an arbitrary way. Formally, we can write

$$U_n(x) = E[\bar{U}_n(x)|\mathbf{X_n}] ,$$

where $\mathbf{X_n}$ is the information containing the original sample but without regard to their order. Applying the inequality $[E(Y)]^2 \le E(Y^2)$ (conditionally) yields

$$E[U_n^2(x)] = E\{E[\bar{U}_n(x)|\mathbf{X_n}]\}^2 \le \{E[\bar{U}_n^2(x)|\mathbf{X_n}]\} = E[\bar{U}_n^2(x)] .$$

Thus, $Var[U_n(x)] \to 0$ and (i) follows.

To prove (ii), given any subsequence $\{n_k\}$, one can extract a further subsequence $\{n_{k_j}\}$ so that $L_{n_{k_j}}(x) \to J(x, P)$ almost surely. Therefore, $L_{n_{k_j}}(x) \to J(x, P)$ almost surely for all x in some countable dense set of the real line. So, $L_{n_{k_j}}$ tends weakly to $J(x, P)$ and this convergence is uniform by Polya's Theorem. Hence, the result (ii) holds.

To prove (iii), if $J(\cdot, P)$ is also assumed strictly increasing at $c(1-\alpha, P)$, then

$$c_{n,b}(1-\alpha) \xrightarrow{P} c(1-\alpha, P)$$

by Problem 11.30, which is a generalization of Lemma 11.2.1. The limiting coverage probability now follows from Slutsky's Theorem. To complete the proof without the strictly increasing assumption, see Problem 18.6. ■

The assumptions $b/n \to 0$ and $b \to \infty$ need not imply $\tau_b/\tau_n \to 0$. For example, in the unusual case $\tau_n = \log(n)$, if $b = n^{\gamma}$ and $\gamma > 0$, the assumption $\tau_b/\tau_n \to 0$ is not satisfied. In fact, a slight modification of the method is consistent without assuming $\tau_b/\tau_n \to 0$; see Politis et al. (1999), Corollary 2.2.1. In regular cases, $\tau_n = n^{1/2}$, and the assumptions on b simplify to $b/n \to 0$ and $b \to \infty$.

The assumptions on b are as weak as possible under the weak assumptions of the theorem. However, in some cases, the choice $b = O(n)$ yields similar results; this occurs in Wu (1990), where the statistic is approximately linear with an asymptotic normal distribution and $\tau_n = n^{1/2}$. This choice will not work in general; see Example 18.7.2.

Assumption 18.7.1 is satisfied in numerous examples, including all previous examples considered by the bootstrap.

18.7.2 Comparison with the Bootstrap

The usual bootstrap approximation to $J_n(x, P)$ is $J_n(x, \hat{Q}_n)$, where $\hat{Q}_n$ is some estimate of P. In many nonparametric i.i.d. situations, $\hat{Q}_n$ is taken to be the empirical distribution of the sample $X_1, \ldots, X_n$. In Section 18.3, we proved results like (18.49) and (18.50) with $L_{n,b}(x)$ replaced by $J_n(x, \hat{Q}_n)$. While the consistency of the bootstrap requires arguments specific to the problem at hand, the consistency of subsampling holds quite generally.

To elaborate a little further, we proved bootstrap limit results in the following manner. For some choice of metric (or pseudo-metric) d on the space of probability measures, it must be known that $d(P_n, P) \to 0$ implies $J_n(P_n)$ converges weakly to $J(P)$. That is, Assumption 18.7.1 must be strengthened so that the convergence of $J_n(P)$ to $J(P)$ is suitably locally uniform in P. In addition, the estimator $\hat{Q}_n$ must then be known to satisfy $d(\hat{Q}_n, P) \to 0$ almost surely or in probability under P. In contrast, no such strengthening of Assumption 18.7.1 is required in Theorem 18.7.1. In the known counterexamples to the bootstrap, it is precisely a certain lack of uniformity in convergence which leads to failure of the bootstrap.

In some special cases, it has been realized that a sample size trick can often remedy the inconsistency of the bootstrap. To describe how, focus on the case where $\hat{Q}_n$ is the empirical measure, denoted by $\hat{P}_n$. Rather than approximating $J_n(P)$ by $J_n(\hat{P}_n)$, the suggestion is to approximate $J_n(P)$ by $J_b(\hat{P}_n)$ for some b which usually satisfies $b/n \to 0$ and $b \to \infty$. The resulting estimator $J_b(x, \hat{P}_n)$ is obviously quite similar to our $L_{n,b}(x)$ given in (2.1). In words, $J_b(x, \hat{P}_n)$ is the bootstrap approximation defined by the distribution (conditional on the data) of $\tau_b[\hat{\theta}_b(X_1^*, \ldots, X_b^*) - \hat{\theta}_n]$, where $X_1^*, \ldots, X_b^*$ are chosen with replacement from $X_1, \ldots, X_n$. In contrast, $L_{n,b}(x)$ is the distribution (conditional on the data) of $\tau_b[\hat{\theta}_b(Y_1^*, \ldots, Y_b^*) - \hat{\theta}_n)]$, where $Y_1^*, \ldots, Y_b^*$

are chosen *without* replacement from $X_1, \ldots, X_n$. Clearly, these two approaches must be similar if b is so small that sampling with and without replacement are essentially the same. Indeed, if one resamples b numbers (or indices) from the set $\{1, \ldots, n\}$, then the chance that none of the indices is duplicated is $\Pi_{i=1}^{b-1}(1 - \frac{i}{n})$. This probability tends to 0 if $b^2/n \to 0$. (To see why, take logs and do a Taylor expansion analysis.) Hence, the following is true.

Corollary 18.7.1 *Under the further assumption that $b^2/n \to 0$, parts (i)–(iii) of Theorem 18.7.1 remain valid if $L_{n,b}(x)$ is replaced by the bootstrap approximation $J_b(x, \hat{P}_n)$.*

The bootstrap approximation with smaller resample size, $J_b(\hat{P}_n)$, is further studied in Bickel, Götze, and van Zwet (1997). In spite of the Corollary, we point out that $L_{n,b}$ is more generally valid. Indeed, without the assumption $b^2/n \to 0$, $J_b(x, \hat{P}_n)$ can be inconsistent. To see why, let P be any distribution on the real line with a density (with respect to Lebesgue measure). Consider any statistic $\hat{\theta}_n$, τ_n, and $\theta(P)$ satisfying Assumption 18.7.1. Even the sample mean will work here. Now, modify $\hat{\theta}_n$ to $\tilde{\theta}_n$ so that the statistic $\tilde{\theta}_n(X_1, \ldots, X_n)$ completely misbehaves if any pair of the observations $X_1, \ldots, X_n$ are identical. The bootstrap approximation to the distribution of $\tilde{\theta}_n$ must then misbehave as well unless $b^2/n \to 0$, while the consistency of $L_{n,b}$ remains intact.

The above example, though artificial, was designed to illustrate a point. We now consider some further examples.

Example 18.7.1 (**U-statistics of Degree 2**) Let $X_1, \ldots, X_n$ be i.i.d. on the line with c.d.f. F. Denote by $\hat{F}_n$ the empirical distribution of the data. Let

$$\theta(F) = \int \int \omega(x, y) dF(x) dF(y)$$

and assume $\omega(x, y) = \omega(y, x)$. Assume

$$\int \omega^2(x, y) dF(x) dF(y) < \infty .$$

Set $\tau_n = n^{1/2}$ and

$$\hat{\theta}_n = \sum_{i<j} \omega(X_i, X_j) / \binom{n}{2} .$$

Then, by Theorem 12.3.2, $J_n(F)$ converges weakly to $J(F)$, the normal distribution with mean 0 and variance given by

$$v^2(F) = 4 \left\{ \int [\omega(x, y) dF(y)]^2 dF(x) - \theta^2(F) \right\} .$$

Hence, Assumption 18.7.1 holds. However, in order for the bootstrap to succeed, the additional condition $\int \omega^2(x, x) dF(x) < \infty$ is required. Bickel and Freedman

(1981) give a counterexample to show the inconsistency of the bootstrap without this additional condition.

Interestingly, the bootstrap may fail even if $\int \omega^2(x,x)dF(x) < \infty$, stemming from the possibility that $v^2(F) = 0$. (Otherwise, Bickel and Freedman's argument justifies the bootstrap.) As an example, let $w(x, y) = xy$. In this case,

$$\theta(\hat{F}_n) = \bar{X}_n^2 - S_n^2/n ,$$

where S_n^2 is the usual unbiased sample variance. If $\theta(F) = 0$, then $v(F) = 0$. Then, $n[\theta(\hat{F}_n) - \theta(F)]$ converges weakly to $\sigma^2(F)(Z^2 - 1)$, where Z denotes a standard normal random variable and $\sigma^2(F)$ denotes the variance of F. However, it is easy to see that the bootstrap approximation to the distribution of $n[\theta(\hat{F}_n) - \theta(F)]$ has a representation $\sigma^2(F)Z^2 + 2Z\sigma(F)n^{1/2}\bar{X}_n$. Thus, failure of the bootstrap follows.

In the context of U-statistics, the possibility of using a reduced sample size in the resampling has been considered in Bretagnolle (1983); an alternative correction is given by Arcones (1991). ■

Example 18.7.2 (**Extreme Order Statistic**) The following counterexample is taken from Bickel and Freedman (1981). If $X_1, \ldots, X_n$ are i.i.d. according to a uniform distribution on $(0, \theta)$, let $X_{(n)}$ be the maximum order statistic. Then, $n[X_{(n)} - \theta]$ has a limit distribution given by the distribution of $-\theta X$, where X is exponential with mean one. Hence, Assumption 18.7.1 is satisfied here. However, the usual bootstrap fails. To see why, let $X_1^*, \ldots, X_n^*$ be n observations sampled from the data with replacement, and let $X_{(n)}^*$ be the maximum of the bootstrap sample. The bootstrap approximation to the distribution of $n[X_{(n)} - \theta]$ is the distribution of $n[X_{(n)}^* - X_{(n)}]$, conditional on $X_1, \ldots, X_n$. But, the probability mass at 0 for this bootstrap distribution is the probability that $X_{(n)}^* = X_{(n)}$, which occurs with probability

$$1 - (1 - \frac{1}{n})^n \to 1 - \exp(1) .$$

However, the true limiting distribution is continuous. Note in Theorem 18.7.1 that the conditions on b (with $\tau_n = n$) reduce to $b/n \to 0$ and $b \to \infty$. In this example, at least, it is clear that we cannot assume $b/n \to c$, where $c > 0$. Indeed, $L_{n,b}(x)$ places mass b/n at 0. Thus, while it is sometimes true that, under further conditions such as Wu (1990) assumes, we can take b to be of the same order as n, this example makes it clear that we cannot in general weaken our assumptions on b without imposing further structure. ■

Example 18.7.3 (**Superefficient Estimator**) Assume $X_1, \ldots, X_n$ are i.i.d. according the normal distribution with mean $\theta(P)$ and variance one. Fix $c > 0$. Let $\hat{\theta}_n = c\bar{X}_n$ if $|\bar{X}_n| \leq n^{-1/4}$ and $\hat{\theta}_n = \bar{X}_n$ otherwise. The resulting estimator is known as Hodges' superefficient estimator; see Lehmann and Casella (1998), p. 440 and Problem 14.70. It is easily checked that $n^{1/2}(\hat{\theta}_n - \theta(P))$ has a limit distribution for every θ, so the conditions for Theorem 18.7.1 remain applicable. However, Beran

(1984) showed that the distribution of $n^{1/2}(\hat{\theta}_n - \theta(P))$ cannot be bootstrapped, even if one is willing to apply a parametric bootstrap! ■

We have claimed that subsampling is superior to the bootstrap in a first-order asymptotic sense, since it is more generally valid. However, in many typical situations, the bootstrap is far superior and has some compelling second-order asymptotic properties. Some of these were studied in Section 18.4; also see Hall (1992). In nice situations, such as when the statistic or root is a smooth function of sample means, a bootstrap approach is often very satisfactory. In other situations, especially those where it is not known that the bootstrap works even in a first-order asymptotic sense, subsampling is preferable. Still, in other situations (such as the mean in the infinite variance case), the bootstrap may work, but only with a reduced sample size. The issue becomes whether to sample with or without replacement (as well as the choice of resample size). Although this question is not yet answered unequivocally, some preliminary evidence in Bickel et al. (1997) suggests that the bootstrap approximation $J_b(x, \hat{P}_n)$ might be more accurate; more details on the issue of higher order accuracy of the subsampling approximation $L_{n,b}(x)$ are given in Chapter 10 of Politis et al. (1999).

Because $\binom{n}{b}$ can be large, $L_{n,b}$ may be difficult to compute. Instead, an approximation may be employed. For example, let $I_1, \ldots I_B$ be chosen randomly with or without replacement from $\{1, 2, \ldots, N_n\}$. Then, $L_{n,b}(x)$ may be approximated by

$$\hat{L}_{n,b}(x) = \frac{1}{B} \sum_{i=1}^{B} I\{\tau_b(\hat{\theta}_{n,b,I_i} - \hat{\theta}_n) \le x\}. \tag{18.52}$$

Corollary 18.7.2 *Under the assumptions of Theorem 18.7.1 and the assumption $B \to \infty$ as $n \to \infty$, the results of Theorem 18.7.1 are valid if $L_{n,b}(x)$ is replaced by $\hat{L}_{n,b}(x)$.*

PROOF. If the I_i are sampled with replacement, $\sup_x |\hat{L}_{n,b}(x) - L_{n,b}(x)| \to 0$ in probability by the Dvoretzky, Kiefer, Wolfowitz inequality. This result is also true in the case the I_i are sampled without replacement; apply Proposition 4.1 of Romano (1989b). ■

An alternative approach, which also requires fewer computations, is the following. Rather than employing all $\binom{n}{b}$ subsamples of size b from $X_1, \ldots, X_n$, just use the $n - b + 1$ subsamples of size b of the form $\{X_i, X_{i+1}, \ldots, X_{i+b-1}\}$. Notice that the ordering of the data is fixed and retained in the subsamples. Indeed, this is the approach that is applied for time series data; see Chapter 3 of Politis et al. (1999), where consistency results in data-dependent situations are given. Even when the i.i.d. assumption seems reasonable, this approach may be desirable to ensure robustness against possible serial correlation. Most inferential procedures based on i.i.d. models are simply not valid (i.e., not even first-order accurate) if the independence assumption is violated, so it seems worthwhile to account for possible dependencies in the data if we do not sacrifice too much in efficiency.

18.7.3 Hypothesis Testing

In this section, we consider the use of subsampling for the construction of hypothesis tests. As before, $X_1, \dots, X_n$ is a sample of n independent and identically distributed observations taking values in a sample space S. The common unknown distribution generating the data is denoted by P. This unknown law P is assumed to belong to a certain class of laws $\mathbf{P}$. The null hypothesis H asserts $P \in \mathbf{P_0}$, and the alternative hypothesis K is $P \in \mathbf{P_1}$, where $\mathbf{P_i} \subset \mathbf{P}$ and $\mathbf{P_0} \bigcup \mathbf{P_1} = \mathbf{P}$.

The goal is to construct an asymptotically valid test based on a given test statistic,

$$T_n = \tau_n t_n(X_1, \dots, X_n) ,$$

where, as before, τ_n is a fixed nonrandom normalizing sequence. Let

$$G_n(x, P) = P\{\tau_n t_n(X_1, \dots, X_n) \le x\} .$$

We will be assuming that $G_n(\cdot, P)$ converges in distribution, at least for $P \in \mathbf{P_0}$. Of course, this would imply (as long as $\tau_n \to \infty$) that $t_n(X_1, \dots, X_n) \to 0$ in probability for $P \in \mathbf{P_0}$. Naturally, t_n should somehow be designed to distinguish between the competing hypotheses. The theorem we will present will assume t_n is constructed to satisfy the following: $t_n(X_1, \dots, X_n) \to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P) = 0$ if $P \in \mathbf{P_0}$ and $t(P) > 0$ if $P \in \mathbf{P_1}$. This assumption easily holds in typical examples.

To describe the test construction, as in Section 18.7.1, let $Y_1, \dots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1, \dots, X_n\}$, ordered in any fashion. Let $t_{n,b,i}$ be equal to the statistic t_b evaluated at the data set Y_i. The sampling distribution of T_n is then approximated by

$$\hat{G}_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} I\{\tau_b t_{n,b,i} \le x\} . \tag{18.53}$$

Using this estimated sampling distribution, the critical value for the test is obtained as the $1 - \alpha$ quantile of $\hat{G}_{n,b}(\cdot)$; specifically, define

$$g_{n,b}(1 - \alpha) = \inf\{x : \hat{G}_{n,b}(x) \ge 1 - \alpha\} . \tag{18.54}$$

Finally, the nominal level α test rejects H if and only if $T_n > g_{n,b}(1 - \alpha)$.

The following theorem gives the asymptotic behavior of this procedure, showing the test is pointwise consistent in level and pointwise consistent in power. In addition, an expression for the limiting power of the test is obtained under a sequence of alternatives contiguous to a distribution in the null hypothesis.

Theorem 18.7.2 *Assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.*

(i) Assume, for $P \in \mathbf{P_0}$, $G_n(P)$ converges weakly to a continuous limit law $G(P)$, whose corresponding cumulative distribution function is $G(\cdot, P)$ and whose

$1-\alpha$ quantile is $g(1-\alpha, P)$. If $G(\cdot, P)$ is continuous at $g(1-\alpha, P)$ and $P \in \mathbf{P_0}$, then

$$g_{n,b}(1-\alpha) \to g(1-\alpha, P) \textit{ in probability}$$

and

$$P\{T_n > g_{n,b}(1-\alpha)\} \to \alpha \textit{ as } n \to \infty.$$

(ii) Assume the test statistic is constructed so that $t_n(X_1, \dots, X_n) \to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P) = 0$ if $P \in \mathbf{P_0}$ and $t(P) > 0$ if $P \in \mathbf{P_1}$. Assume $\liminf_n(\tau_n/\tau_b) > 1$. Then, if $P \in \mathbf{P_1}$, the rejection probability satisfies

$$P\{T_n > g_{n,b}(1-\alpha)\} \to 1 \textit{ as } n \to \infty.$$

(iii) Suppose P_n is a sequence of alternatives such that, for some $P_0 \in \mathbf{P_0}$, $\{P_n^n\}$ is contiguous to $\{P_0^n\}$. Then,

$$g_{n,b}(1-\alpha) \to g(1-\alpha, P_0) \textit{ in } P_n^n\textit{-probability.}$$

Hence, if T_n converges in distribution to T under P_n and $G(\cdot, P_0)$ is continuous at $g(1-\alpha, P_0)$, then

$$P_n^n\{T_n > g_{n,b}(1-\alpha)\} \to Prob\{T > g(1-\alpha, P_0)\}.$$

The proof is similar to that of Theorem 18.7.1 (Problem 18.45).

Example 18.7.4 Consider the special case of testing a real-valued parameter. Specifically, suppose $\theta(\cdot)$ is a real-valued function from $\mathbf{P}$ to the real line. The null hypothesis is specified by $\mathbf{P_0} = \{P : \theta(P) = \theta_0\}$. Assume the alternative is one sided and is specified by $\{P : \theta(P) > \theta_0\}$. Suppose we simply take

$$t_n(X_1, \dots, X_n) = \hat{\theta}_n(X_1, \dots, X_n) - \theta_0 .$$

If $\hat{\theta}_n$ is a consistent estimator of $\theta(P)$, then the hypothesis on t_n in part (ii) of the theorem is satisfied (just take the absolute value of t_n for a two-sided alternative). Thus, the hypothesis on t_n in part (ii) of the theorem boils down to verifying a consistency property and is rather weak, though this assumption can in fact be weakened further. The convergence hypothesis of part (i) is satisfied by typical test statistics; in regular situations, $\tau_n = n^{1/2}$. ■

The interpretation of part (iii) of the theorem is the following. Suppose, instead of using the subsampling construction, one could use the test that rejects when $T_n > g_n(1-\alpha, P)$, where $g_n(1-\alpha, P)$ is the exact $1-\alpha$ quantile of the true sampling distribution $G_n(\cdot, P)$. Of course, this test is not available in general because P is unknown and so is $g_n(1-\alpha, P)$. Then, the asymptotic power of the subsampling test against a sequence of contiguous alternatives $\{P_n\}$ to P with P in $\mathbf{P_0}$ is the same as the asymptotic power of this fictitious test against the same sequence of alternatives. Hence, to the order considered, there is no loss in efficiency in terms of power.

Example 18.7.5 (**Moment Inequalities Using Subsampling**) Reconsider the moment inequality testing problem in Example 18.5.1, where the problem is to test all components of a mean vector are less than or equal to zero. Let $\tau_n = \sqrt{n}$ and

$$t_n(X_1, \ldots, X_n) = \max_{1 \le i \le k} \bar{X}_{n,i} \ .$$

The subsampling distribution is then defined as in (18.53). Theorem 18.7.2 applies and the subsampling test controls Type 1 error asymptotically. (In fact, it controls Type 1 error uniformly over a large collection of underlying distributions; see Romano and Shaikh (2008, 2012).) Looking at the local asymptotic power properties of this test, suppose $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,k})^\top$ lies on the boundary of the parameter space, so that $\theta_{0,i} \le 0$ for all i and equal to 0 for some i. Also, let

$$I = \{i : \theta_i(P) = 0\}$$

and $h = (h_1, \ldots, h_k)^\top$. To keep it simple, suppose $P = P_\theta$ is multivariate normal with mean θ and covariance matrix Σ. Then, under $\theta_0 + hn^{-1/2}$,

$$T_n = \sqrt{n} t_n \xrightarrow{d} \max_{i \in I}(Z_i + h_i) \ , \tag{18.55}$$

where $(Z_1, \ldots, Z_k)^\top$ is multivariate normal with mean 0 and covariance matrix Σ. Under such a sequence, the subsampling distribution $\hat{G}_{n,b}(x)$ satisfies, for any x,

$$\hat{G}_{n,b}(x) \xrightarrow{P} P\{\max_{i \in I} Z_i \le x\} \ . \tag{18.56}$$

Therefore, if $d_{I,1-\alpha}$ denotes the $1-\alpha$ quantile of $\max_{i \in I} Z_i$, then the subsampling quantile $g_{n,b}(1-\alpha)$ satisfies

$$g_{n,b}(1-\alpha) \xrightarrow{P} d_{I,1-\alpha} \ .$$

Hence, the limiting power against $\theta_0 + hn^{-1/2}$ of the subsampling test can be expressed as (Problem 18.48)

$$P\{\max_{i \in I}(Z_i + h_i) > d_{I,1-\alpha}\} \ . \tag{18.57}$$

In particular, this limiting power is greater than it would be if $d_{I,1-\alpha}$ were replaced by $d_{I_0,1-\alpha}$, where $I_0 = \{1, \ldots, k\}$. In other words, subsampling implicitly is applying a moment selection procedure. Compare with the bootstrap in Problems 14.68 and 18.30. ■

18.8 Problems

Section 18.2

Problem 18.1 Assume $X_1, \ldots, X_n$ are i.i.d. according to a location-scale model with distribution of the form $F[(x-\theta)/\sigma]$, where F is known, θ is a location parameter, and σ is a scale parameter. Suppose $\hat{\theta}_n$ is a location and scale equivariant estimator and $\hat{\sigma}_n$ is a location invariant, scale equivariant estimator. Then, show that the roots $[\hat{\theta}_n - \theta]/\hat{\sigma}_n$ and $\hat{\sigma}_n/\sigma$ are pivots.

Problem 18.2 Let $X = (X_1, \ldots, X_n)^\top$ and consider the linear model

$$X_i = \sum_{j=1}^{s} a_{i,j}\beta_j + \sigma\epsilon_i \ ,$$

where the ϵ_i are i.i.d. F, where F has mean 0 and variance 1. Here, the $a_{i,j}$ are known, $\beta = (\beta_1, \ldots, \beta_s)^\top$ and σ are unknown. Let A be the $n \times s$ matrix with (i, j) entry $a_{i,j}$ and assume A has rank s. As in Section 13.2.3, let $\hat{\beta}_n = (A^\top A)^{-1}A^\top X$ be the least squares estimate of β. Consider the test statistic

$$T_n = \frac{(n-s)(\hat{\beta}_n - \beta)(A^\top A)(\hat{\beta}_n - \beta)}{sS_n^2} \ ,$$

where $S_n^2 = (X - A\hat{\beta}_n)^\top(X - A\hat{\beta}_n)/(n-s)$. Is T_n a pivot when F is known?

Section 18.3

Problem 18.3 Suppose the convergence (18.4) only holds in probability and that $J(\cdot, P)$ is continuous and strictly increasing at $J^{-1}(1-\alpha, P)$. Show that (18.5) holds in probability. Then, show that (18.6) still holds.

Problem 18.4 In Theorem 18.3.1, one cannot deduce the uniform convergence result (18.4) without the assumption that the limit law $J(P)$ is continuous. Show that, without the continuity assumption for $J(P)$,

$$\rho_L(J_n(\hat{P}_n), J_n(P)) \to 0$$

with probability one, where ρ_L is the Lévy metric defined in Definition 11.2.3.

Problem 18.5 In Theorem 18.3.3 (i), show that the assumption that $\theta(F_n) \to \theta(F)$ actually follows from the other assumptions.

Problem 18.6 Under the assumptions of Theorem 18.3.1 but without the assumption that $J(\cdot, P)$ is strictly increasing, show that the conclusion (18.6) still holds. [*Hint:* See Problems 11.26 and 11.31.]

Problem 18.7 Reprove Theorem 18.3.3(ii) under the assumption $E(|X_i|^3) < \infty$ by using the Berry–Esseen Theorem.

Problem 18.8 Prove the following extension of Theorem 18.3.3 holds. Let $\mathbf{D_F}$ be the set of sequences $\{F_n\}$ such that F_n converges weakly to a distribution G and $\sigma^2(F_n) \to \sigma^2(G) = \sigma^2(F)$. Then, Theorem (18.3.3) holds with $\mathbf{C}_F$ replaced by $\mathbf{D_F}$. (Actually, one really only needs to define $\mathbf{D_F}$ so that and sequence $\{F_n\}$ is tight and any weakly convergent subsequence of $\{F_n\}$ has the above property.) Thus, the possible choices for the resampling distribution are quite large in the sense that the bootstrap approximation $J_n(\hat{G}_n)$ can be consistent even if $\hat{G}_n$ is not at all close to F. For example, the choice where $\hat{G}_n$ is normal with mean $\bar{X}_n$ and variance equal to a consistent estimate of the sample variance results in consistency. Therefore, the normal approximation can in fact be viewed as a bootstrap procedure with a perverse choice of resampling distribution. Show the bootstrap can be inconsistent if $\sigma^2(G) \neq \sigma^2(F)$.

Problem 18.9 Assume $X_1, \ldots, X_n$ are i.i.d. with c.d.f. F, mean $\mu(F)$ and variance $\sigma^2(F) < \infty$. Let $\bar{X}_n = \mu(\hat{F}_n)$, where $\hat{F}_n$ is the empirical c.d.f. Conditional on $\hat{F}_n$, let $X_1^*, \ldots, X_n^*$ be i.i.d. according to $\hat{F}_n$, with sample mean $\bar{X}_n^*$. Find the (unconditional) joint limiting distribution of

$$n^{1/2}[\bar{X}_n^* - \bar{X}_n, \bar{X}_n - \mu(F)] .$$

Problem 18.10 Under the setup of Problem 18.9, the problem now is to construct a confidence interval for $\mu(F)$, but now it is known and assumed that $\mu(F) \geq 0$. Inference is based on the estimator $\hat{\mu}_n = \max(\bar{X}_n, 0)$. Consider the root $R_n(X_1, \ldots, X_n, \mu(F)) = n^{1/2}[\hat{\mu}_n - \mu(F)]$, with distribution $J_n(F)$. Investigate bootstrap consistency. Separate out cases by $\mu(F) > 0$ and $\mu(F) = 0$. *Hint: In the case $\mu(F) = 0$, first find the limiting behavior of $J_n(F)$. For any $c > 0$, if $n^{1/2}\bar{X}_n < -c$, show that the bootstrap distribution is dominated in the limit by that of* $\sigma(F)\max(Z - c, 0)$*, where $Z \sim N(0, 1)$. Use the almost sure representation theorem to argue that the bootstrap fails, at least along a subsequence.*

Problem 18.11 In the case that $\theta(P)$ is real-valued, Efron initially proposed the following construction, called the bootstrap *percentile* method. Let $\hat{\theta}_n$ be an estimator of $\theta(P)$, and let $\tilde{J}_n(P)$ be the distribution of $\hat{\theta}_n$ under P. Then, Efron's two-sided percentile interval of nominal level $1 - \alpha$ takes the form

$$[\tilde{J}_n^{-1}(\frac{\alpha}{2}, \hat{P}_n), \tilde{J}_n^{-1}(1 - \frac{\alpha}{2}, \hat{P}_n)] . \tag{18.58}$$

Also, consider the root $R_n(X^n, \theta(P)) = n^{1/2}(\hat{\theta}_n - \theta(P))$, with distribution $J_n(P)$. Write (18.58) as a function of $\hat{\theta}_n$ and the quantiles of $J_n(\hat{P}_n)$, assuming $\theta(\hat{P}_n) = \hat{\theta}_n$.

Suppose Theorem 18.3.1 holds for the root R_n, so that $J_n(P)$ converges weakly to $J(P)$. What must be assumed about $J(P)$ so that $P\{\theta(P) \in I_n\} \to 1-\alpha$?

Problem 18.12 Let $\hat{\theta}_n$ be an estimate of a real-valued parameter $\theta(P)$. Suppose there exists an increasing transformation g such that

$$g(\hat{\theta}_n) - g(\theta(P))$$

is a pivot, so that its distribution does not depend on P. Also, assume this distribution is continuous, strictly increasing, and symmetric about zero.
(i) Show that Efron's percentile interval (18.58), which may be constructed without knowledge of g, has exact coverage $1-\alpha$.
(ii) Show that the percentile interval is transformation equivariant. That is, if $\phi = m(\theta)$ is a monotone transformation of θ, then the percentile interval for ϕ is the percentile interval for θ transformed by m when $\hat{\phi}_n$ is taken to be $m(\hat{\theta}_n)$. This holds true for the theoretical percentile interval as well as its approximation due to simulation.
(iii) If the parameter θ only takes values in an interval I and $\hat{\theta}_n$ does as well, then the percentile interval is range-preserving in the sense that the interval is always a subset of I.

Problem 18.13 Suppose $\hat{\theta}_n$ is an estimate of some real-valued parameter $\theta(P)$. Let $H_n(x,\theta)$ denote the c.d.f. of $\hat{\theta}_n$ under θ, with inverse $H_n^{-1}(1-\alpha,\theta)$. The percentile interval lower confidence bound of level $1-\alpha$ is then $H_n^{-1}(\alpha,\hat{\theta}_n)$. Suppose that, for some increasing transformation g, and constants z_0 (called the *bias correction*) and a (called the *acceleration constant*),

$$P\{\frac{g(\hat{\theta}_n) - g(\theta)}{1 + ag(\theta)} + z_0 \le x\} = \Phi(x) , \tag{18.59}$$

where Φ is the standard normal c.d.f.
(i) Letting $\hat{\phi}_n = g(\hat{\theta}_n)$, show that $\hat{\theta}_{n,L}$ given by

$$\hat{\theta}_{n,L} = g^{-1}\left\{\hat{\phi}_n + (z_\alpha + z)(1 + a\hat{\phi}_n)/[1 - a(z_\alpha + z_0)]\right\}$$

is an exact $1-\alpha$ lower confidence bound for θ.
(ii) Because $\hat{\theta}_{n,L}$ requires knowledge of g, let

$$\hat{\theta}_{n,BC_a} = H_n^{-1}(\beta, \hat{\theta}_n) ,$$

where

$$\beta = \Phi(z + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)] .$$

Show that $\hat{\theta}_{n,BC_a} = \hat{\theta}_{n,L}$. [The lower bound $\hat{\theta}_{n,BC_a}$ is called the BC_a lower bound and Efron shows one may take $z = \Phi^{-1}(H_n(\hat{\theta}_n, \hat{\theta}_n))$ and gives methods to estimate a; see Efron and Tibshirani (1993, Chapter 14).]

Problem 18.14 Assume the setup of Problem 18.13 and condition (18.59). Let θ_0 be any value of θ and let $\theta_1 = H_n^{-1}(1-\alpha, \theta_0)$. Let

$$\hat{\theta}_{n,AP} = H_n^{-1}(\beta', \hat{\theta}_n) ,$$

where

$$\beta' = H_n(\theta_0, \theta_1) .$$

Show that $\hat{\theta}_{n,AP}$ is an exact level $1-\alpha$ lower confidence bound for θ. [This is called the *automatic percentile* lower bound of DiCiccio and Romano (1989), and may be computed without knowledge of g, a or z. Its exactness holds under assumptions even weaker than (18.59).]

Problem 18.15 Let $X_1, \ldots, X_{n_X}$ be i.i.d. with distribution F_X, and let $Y_1, \ldots, Y_{n_Y}$ be i.i.d. with distribution F_Y. The two samples are independent. Let $\mu(F)$ denote the mean of a distribution F, and let $\sigma^2(F)$ denote the variance of F. Assume $\sigma^2(F_X)$ and $\sigma^2(F_Y)$ are finite. Suppose we are interested in $\theta = \theta(F_X, F_Y) = \mu(F_X) - \mu(F_Y)$. Construct a bootstrap confidence interval for θ of nominal level $1-\alpha$, and prove that it asymptotically has the correct coverage probability, assuming $\min(n_X, n_Y) \to \infty$.

Problem 18.16 Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli trials with success probability θ.
(i). As explicitly as possible, find a uniformly most accurate upper confidence bound for θ of nominal level $1-\alpha$. State the bound explicitly in the case $X_i = 0$ for every i.
(ii). Describe a bootstrap procedure to obtain an upper confidence bound for θ of nominal level $1-\alpha$. What does it reduce to for the previous data set?
(iii). Let $\hat{B}_{1-\alpha}$ denote your upper bootstrap confidence bound for θ. Then, $P_\theta(\theta \leq \hat{B}_{1-\alpha}) \to 1-\alpha$ as $n \to \infty$. Prove the following.

$$\sup_\theta |P_\theta(\theta \leq \hat{B}_{1-\alpha}) - (1-\alpha)|$$

does not tend to 0 as $n \to \infty$.

Problem 18.17 Let $X_1, \ldots, X_n$ be i.i.d. with c.d.f. F, mean $\mu(F)$ and finite variance $\sigma^2(F)$. Consider the root $R_n = n^{1/2}(\bar{X}_n^2 - \mu^2(F))$ and the bootstrap approximation to its distribution $J_n(\hat{F}_n)$, where $\hat{F}_n$ is the empirical c.d.f. Determine the asymptotic behavior of $J_n(\hat{F}_n)$. *Hint:* Distinguish the cases $\mu(F) = 0$ and $\mu(F) \neq 0$.

Problem 18.18 Prove the remaining details for Theorem 18.3.6. Furthermore, without assuming the differential of f is continuous, one can replace the convergence with probability one results by convergence in probability. (More general results are available in van der Vaart and Wellner (1996).)

Problem 18.19 Let $\epsilon_1, \epsilon_2, \ldots$ be i.i.d. $N(0,1)$. Let $X_i = \mu + \epsilon_i + \beta\epsilon_{i+1}$ with β a fixed nonzero constant. The X_i form a moving-average process studied in Section 13.2.1.
(i) Examine the behavior of the nonparametric bootstrap method for estimating the mean using the root $n^{1/2}(\bar{X}_n - \mu)$ and resampling from the empirical distribution. Show that the coverage probability need not tend to the nominal level under such a moving-average process.
(ii) Suppose $n = bk$ for integers b and k. Consider the following *moving blocks bootstrap* resampling scheme. Let $L_{i,b} = (X_i, X_{i+1}, \ldots, X_{i+b-1})$ be the block of b observations beginning at "time" i. Let $X_1^*, \ldots, X_n^*$ be obtained by randomly choosing with replacement k of the $n - b + 1$ blocks $L_{i,b}$; that is, $X_1^*, \ldots, X_b^*$ are the observations in the first sampled block, $X_{b+1}^*, \ldots, X_{2b}^*$ are the observations from the second sampled block, etc. Then, the distribution of $n^{1/2}[\bar{X}_n - \mu]$ is approximated by the *moving blocks bootstrap* distribution given by the distribution of $n^{1/2}[\bar{X}_n^* - \bar{X}_n]$, where $\bar{X}_n^* = \sum_{i=1}^n X_i^*/n$. If b is fixed, determine the mean and variance of this distribution as $n \to \infty$. Now let $b \to \infty$ as $n \to \infty$. At what rate should $b \to \infty$ so that the mean and variance of the moving blocks distribution tends to the same limiting values as the true mean and variance, at least in probability? [The moving blocks bootstrap was independently discovered by Künsch (1989) and Liu and Singh (1992). The stationary bootstrap of Politis and Romano (1994a) and other methods designed for dependent data are studied in Lahiri (2003).]

Section 18.4

Problem 18.20 Under the assumptions of Theorem 18.4.2, show that, for any $\epsilon > 0$, the expansion (18.20) holds uniformly in $\alpha \in [\epsilon, 1-\epsilon]$.

Problem 18.21 Under the assumptions of Theorem 18.4.1, show that, for any $\epsilon > 0$, the expansion (18.21) holds uniformly in $\alpha \in [\epsilon, 1-\epsilon]$.

Problem 18.22 Suppose Y_n is a sequence of random variables satisfying

$$P\{Y_n \le t\} = g_0(t) + g_1(t)n^{-1/2} + O(n^{-1}) ,$$

uniformly in t, where g_0 and g_1 have uniformly bounded derivatives. If $T_n = O_P(n^{-1})$, then show, for any fixed (nonrandom) sequence t_n,

$$P\{Y_n \le t_n + T_n\} = g_0(t_n) + g_1(t_n)n^{-1/2} + O(n^{-1}) .$$

Problem 18.23 Assuming the expansions in the section hold, show that the two-sided bootstrap interval (18.29) has coverage error of order n^{-1}.

Problem 18.24 Assuming the expansions in the section hold, show that the two-sided bootstrap-t interval (18.35) has coverage error of order n^{-1}.

Problem 18.25 Verify the expansion (18.36) and argue that the resulting interval $I_n(1-\hat{\alpha}_n)$ has coverage error $O(n^{-1})$.

Problem 18.26 In the nonparametric mean setting, determine the one- and two-sided coverage errors of Efron's percentile method described in (18.58).

Problem 18.27 Assume F has infinitely many moments and is absolutely continuous. Under the notation of this section, argue that $n^{1/2}[J_n(t, \hat{F}_n) - J_n(t, F)]$ has an asymptotically normal limiting distribution, as does $n[K_n(t, \hat{F}_n) - K_n(t, F)]$.

Problem 18.28 (i) In a normal location model $N(\mu, \sigma^2)$, consider the root $R_n = n^{1/2}(\bar{X}_n - \mu)$, which is not a pivot. Show that bootstrap calibration, by parametric resampling, produces an exact interval.
(ii) Next, consider the root $n^{1/2}(S_n^2 - \sigma^2)$, where S_n^2 is the usual unbiased estimate of variance. Show that bootstrap calibration, by parametric resampling, produces an exact interval.

Problem 18.29 (i) Show the bootstrap interval (18.3) can be written as

$$\{\theta \in \Theta :\ J_n(R_n(X^n, \theta), \hat{P}_n) \le 1 - \alpha\} \tag{18.60}$$

if, for the purposes of this problem, $J_n(x, P)$ is defined as the left continuous c.d.f.

$$J_n(x, P) = P\{R_n(X^n, \theta(P)) < x\}$$

and $J_n^{-1}(1-\alpha, P)$ is now defined as

$$J_n^{-1}(1-\alpha, P) = \sup\{x :\ J_n(x, P) \le 1 - \alpha\}\ .$$

[*Hint:* If a random variable Y has left continuous c.d.f. $F(x) = P\{Y < x\}$ and $F^{-1}(1-\alpha)$ is the largest $1-\alpha$ quantile of F, then the event $\{X \le F^{-1}(1-\alpha)\}$ is identical to $\{F(X) \le 1-\alpha\}$ for any random variable X (which need not have distribution F). Why?]
(ii) The bootstrap interval (18.60) pretends that

$$R_{n,1}(X^n, \theta(P)) \equiv J_n(R_n(X^n, \theta(P)), \hat{P}_n)$$

has the uniform distribution on (0, 1). Let $J_{n,1}(P)$ be the actual distribution of $R_{n,1}(X^n, \theta(P))$ under P, with left continuous c.d.f. denoted by $J_{n,1}(x, P)$. This results in a new interval with R_n and J_n replaced by $R_{n,1}$ and $J_{n,1}$ in (18.60). Show that the resulting interval is equivalent to bootstrap calibration of the initial interval. [The mapping of R_n into $R_{n,1}$ by estimated c.d.f. of the former is called *prepivoting*. Beran (1987, 1988b) argues that the interval based on $R_{n,1}$ has better coverage properties than the interval based on R_n.]

Section 18.5

Problem 18.30 In Example 18.5.1, assume P is multivariate normal. Fix θ_0 on the boundary of the null hypothesis parameter space, so that $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,k})^\top$ has $\theta_{0,i} \le 0$ for all i and equal to zero for at least one i. Fix $h = (h_1, \ldots, h_k)^\top$ and calculate the limiting power of the bootstrap test against alternatives $\theta_0 + hn^{-1/2}$. Compare with Problem 14.68.

Problem 18.31 Explain why the parametric bootstrap method described in Example 18.5.7 fails if $k > 1$. What happens if $k = 1$? [By sufficiency, you may assume you observe $X = (X_1, \ldots, X_k)^\top$ multivariate normal. In the case $k = 2$, plot the rejection region. Recall Examples 8.7.3 and 14.4.8).]

Problem 18.32 In Example 18.5.2, rather than exact evaluation of $G_n(\cdot, \hat{Q}_n)$, describe a simulation test of H that has exact level α.

Problem 18.33 In Example 18.5.3, why is the parametric bootstrap test exact for the special case of Example 14.4.7?

Problem 18.34 In the Behrens–Fisher problem, show that (18.37) and (18.38) hold.

Problem 18.35 In the Behrens–Fisher problem, verify the bootstrap-t has rejection probability equal to $\alpha + O(n^{-2})$.

Problem 18.36 In the Behrens–Fisher problem, what is the order of error in rejection probability for the likelihood ratio test? What is the order of error in rejection probability if you bootstrap the non-studentized statistic $n^{1/2}(\bar{X}_{n,1} - \bar{X}_{n,2})$.

Problem 18.37 In Example 18.5.5, with resampling from the empirical distribution shifted to have mean 0, what are the errors in rejection for the tests based on T_n and T_n'? How do these tests differ from the corresponding tests obtained through inverting bootstrap confidence bounds?

Problem 18.38 Let $X_1, \ldots, X_n$ be i.i.d. with a distribution P on the real line, and let $\hat{P}_n$ be the empirical distribution function. Find Q that minimizes $\delta_{KL}(\hat{P}_n, Q)$, where δ_{KL} is the Kullback–Leibler divergence defined by (18.39).

Problem 18.39 Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued with c.d.f. F. The problem is to test the null hypothesis that F is $N(\mu, \sigma^2)$ for some (μ, σ^2). Consider the test statistic

$$T_n = n^{1/2} \sup_t |\hat{F}_n(t) - \Phi((t - \bar{X}_n)/\hat{\sigma}_n)| ,$$

where $\hat{F}_n$ is the empirical c.d.f. and $(\bar{X}_n, \hat{\sigma}_n^2)$ is the MLE for (μ, σ^2) assuming normality. Argue that the distribution of T_n does not depend on (μ, σ^2) and describe an exact bootstrap test construction. [Such problems are studied in Romano (1988)].

Section 18.6

Problem 18.40 Show why (18.47) is true.

Problem 18.41 (i) Under the setup of Example 18.6.1, prove that Theorem 18.6.1 applies if studentized statistics are used.
(ii) In addition to the $X_1, \ldots, X_n$, suppose i.i.d. $Y_1, \ldots, Y_{n'}$ are observed, with $Y_i = (Y_{i,1}, \ldots, Y_{i,s})$. The distribution of Y_i need not be that of X_i. Suppose the mean of Y_i is $(\mu'_1, \ldots, \mu'_s)$. Generalize Example 18.6.1 to simultaneously test $H_i : \mu_i = \mu'_i$. Distinguish between two cases, first where the X_is are independent of the Y_js, and next where (X_i, Y_i) are paired (so $n = n'$) and X_i need not be independent of Y_i.

Problem 18.42 Under the setup of Example 18.6.2, provide the details to show that the FWER is asymptotically controlled.

Problem 18.43 Under the setup of Example 18.6.2, suppose that there is also an i.i.d. control sample $X_{0,1}, \ldots, X_{0,n_0}$, independent of the other Xs. Let μ_0 denote the mean of the controls. Now consider testing $H_i : \mu_i = \mu_0$. Describe a method that asymptotically controls the FWER.

Problem 18.44 Under the setup of Example 18.6.2, let F_i denote the distribution of the ith sample. Now, consider $H'_{i,j} : F_i = F_j$ based on the same test statistics. Describe a randomization test that controls the FWER.

Section 18.7

Problem 18.45 Prove Theorem 18.7.2. [*Hint*: For (ii), rather than considering $\hat{G}_{n,b}(x)$, just look at the empirical distribution, $\hat{G}^0_{n,b}$, of the values of $t_{n,b,i}$ (not scaled by τ_b) and show $\hat{G}^0_{n,b}(\cdot)$ converges in distribution to a point mass at $t(P)$.]

Problem 18.46 Prove a general subsampling theorem for two-sample problems. Here, you observe $X_1, \ldots, X_m$ i.i.d. P and independently $Y_1, \ldots, Y_n$ are i.i.d. Q. The problem is to get a confidence interval for $\theta(Q) - \theta(P)$. Assume $\min(m, n) \to \infty$. Describe the method, state a theorem, and prove it.

Problem 18.47 Prove a result for subsampling analogous to Theorem 18.6.1, but that does not require assumption (18.46). [Theorem 18.6.1 applies to testing real-valued parameters; a more general multiple testing procedure based on subsampling is given by Theorem 4.4 of Romano and Wolf (2005a).]

Problem 18.48 In Example 18.7.5, verify (18.55), (18.56), and (18.57).

Problem 18.49 To see how subsampling extends to a dependent time series model, assume $X_1, \ldots, X_n$ are sampled from a stationary time series model that is m-dependent. [Stationarity means the distribution of the $X_1, X_2, \ldots$ is the same as that of $X_t, X_{t+1}, \dots$ for any t. The process is m-dependent if, for any t and m, $(X_1, \dots, X_t)$ and $(X_{t+m+1}, X_{t+m+2}, \dots)$ are independent; that is, observations separated in time by more than m units are independent.] Suppose the sum in the definition (18.48) of $L_{n,b}$ extends only over the $n - b + 1$ subsamples of size b of the form $(X_i, X_{i+1}, \dots, X_{i+b-1})$; call the resulting estimate $\tilde{L}_{n,b}$. Under the assumption of stationarity and m-dependence, prove a theorem analogous to Theorem 18.7.1. Then, extend the argument to strong mixing sequences, which were discussed in Section 12.4.

18.9 Notes

The bootstrap was discovered by Efron (1979), who coined the name. Much of the theoretical foundations of the bootstrap are laid out in Bickel and Freedman (1981) and Singh (1981). The development in Section 18.3 is based on Beran (1984). The use of Edgeworth expansions to study the bootstrap was initiated in Singh (1981) and Babu and Singh (1983), and is used prominently in Hall (1992). There have since been hundreds of papers on the bootstrap, as well as several book length treatments, including Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), Davison and Hinkley (1997) and Lahiri (2003). Comparisons of bootstrap and randomization tests are made in Romano (1989b) and Janssen and Pauls (2003b). Westfall and Young (1993) and van der Lann et al. (2004) apply resampling to multiple testing problems. Theorem 18.6.1 is based on Romano and Wolf (2005a). Efficient computation of adjusted p-values for resampling based stepdown multiple testing methods are discussed in Romano and Wolf (2016) and Clarke, Romano and Wolf (2020). Simultaneous bootstrap confidence intervals for differences as described in Example 18.6.2, form a basis for inference for ranks of populations; see Mogstad et al. (2020). Bootstrap results in high-dimensional problems are developed in Chernozhukov et al. (2017) and Xue and Yao (2020).

The method of empirical likelihood referred to in Example 18.5.5 is fully treated in Owen (2001). Similar to parametric models, the method of empirical likelihood can be improved through a Bartlett correction, yielding two-sided tests with error in rejection probability of $O(n^{-2})$; see DiCiccio et al. (1991). Alternatively, rather than using the asymptotic Chi-squared distribution to get critical values, a direct bootstrap approach resamples from $\hat{Q}_n$. Higher order properties of such procedures are considered in DiCiccio and Romano (1990).

The roots of subsampling can be traced to Quenouille's (1949) and Tukey's (1958a) jackknife. Hartigan (1969) and Wu (1990) used subsamples to construct confidence intervals, but in a very limited setting. A general theory for using subsampling to approximate a sampling distribution is presented in Politis and Romano (1994b), including i.i.d. and data-dependent settings. Multi-samples are treated in Politis and Romano (2008, 2010) and McMurry et al. (2012). A full treatment with numerous earlier references is given by Politis et al. (1999). Romano and Shaikh (2012) discuss the uniform asymptotic validity of both the bootstrap and subsampling.