23

Bootstrap

This chapter investigates the asymptotic properties of bootstrap estimators for distributions and confidence intervals. The consistency of the bootstrap for the sample mean implies the consistency for many other statistics by the delta method. A similar result is valid with the empirical process.

23.1 Introduction

In most estimation problems it is important to give an indication of the precision of a given estimate. A simple method is to provide an estimate of the bias and variance of the estimator; more accurate is a confidence interval for the parameter. In this chapter we concentrate on bootstrap confidence intervals and, more generally, discuss the bootstrap as a method of estimating the distribution of a given statistic.

Let $\hat{\theta}$ be an estimator of some parameter θ attached to the distribution P of the observations. The distribution of the difference $\hat{\theta} - \theta$ contains all the information needed for assessing the precision of $\hat{\theta}$. In particular, if ξ_α is the upper α-quantile of the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$, then

$$\mathrm{P}(\hat{\theta} - \xi_\beta \hat{\sigma} \leq \theta \leq \hat{\theta} - \xi_{1-\alpha} \hat{\sigma} \mid P) \geq 1 - \beta - \alpha.$$

Here $\hat{\sigma}$ may be arbitrary, but it is typically an estimate of the standard deviation of $\hat{\theta}$. It follows that the interval $[\hat{\theta} - \xi_\beta \hat{\sigma}, \hat{\theta} - \xi_{1-\alpha} \hat{\sigma}]$ is a confidence interval of level $1 - \beta - \alpha$. Unfortunately, in most situations the quantiles and the distribution of $\hat{\theta} - \theta$ depend on the unknown distribution P of the observations and cannot be used to assess the performance of $\hat{\theta}$. They must be replaced by estimators.

If the sequence $(\hat{\theta} - \theta)/\hat{\sigma}$ tends in distribution to a standard normal variable, then the normal $N(0, \hat{\sigma}^2)$-distribution can be used as an estimator of the distribution of $\hat{\theta} - \theta$, and we can substitute the standard normal quantiles z_α for the quantiles ξ_α. The weak convergence implies that the interval $[\hat{\theta} - z_\beta \hat{\sigma}, \hat{\theta} - z_{1-\alpha} \hat{\sigma}]$ is a confidence interval of asymptotic level $1 - \alpha - \beta$.

Bootstrap procedures yield an alternative. They are based on an estimate $\hat{P}$ of the underlying distribution P of the observations. The distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$ under P can, in principle, be written as a function of P. The bootstrap estimator for this distribution is the "plug-in" estimator obtained by substituting $\hat{P}$ for P in this function. Bootstrap estimators

for quantiles, and next confidence intervals, are obtained from the bootstrap estimator for the distribution.

The following type of notation is customary. Let $\hat{\theta}^*$ and $\hat{\sigma}^*$ be computed from (hypothetic) observations obtained according to $\hat{P}$ in the same way $\hat{\theta}$ and $\hat{\sigma}$ are computed from the true observations with distribution P. If $\hat{\theta}$ is related to $\hat{P}$ in the same way θ is related to P, then the bootstrap estimator for the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$ under P is the distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ under $\hat{P}$. The latter is evaluated given the original observations, that is, for a fixed realization of $\hat{P}$.

A bootstrap estimator for a quantile ξ_α of $(\hat{\theta} - \theta)/\hat{\sigma}$ is a quantile of the distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ under $\hat{P}$. This is the smallest value $x = \hat{\xi}_\alpha$ that satisfies the inequality

$$\mathrm{P}\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*} \le x \mid \hat{P}\right) \ge 1 - \alpha. \tag{23.1}$$

The notation $\mathrm{P}(\cdot \mid \hat{P})$ indicates that the distribution of $(\hat{\theta}^*, \hat{\sigma}^*)$ must be evaluated assuming that the observations are sampled according to $\hat{P}$ *given* the original observations. In particular, in the preceding display $\hat{\theta}$ is to be considered nonrandom. The left side of the preceding display is a function of the original observations, whence the same is true for $\hat{\xi}_\alpha$.

If $\hat{P}$ is close to the true underlying distribution P, then the bootstrap quantiles should be close to the true quantiles, whence it should be true that

$$\mathrm{P}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}} \le \hat{\xi}_\alpha \mid P\right) \approx 1 - \alpha.$$

In this chapter we show that this approximation is valid in an asymptotic sense: The probability on the left converges to $1 - \alpha$ as the number of observations tends to infinity. Thus, the bootstrap confidence interval

$$[\hat{\theta} - \hat{\xi}_\beta \hat{\sigma}, \hat{\theta} - \hat{\xi}_{1-\alpha}\hat{\sigma}] = \left\{\theta : \hat{\xi}_{1-\alpha} \le \frac{\hat{\theta} - \theta}{\hat{\sigma}} \le \hat{\xi}_\beta\right\}$$

possesses asymptotic confidence level $1 - \alpha - \beta$.

The statistic $\hat{\sigma}$ is typically chosen equal to an estimator of the (asymptotic) standard deviation of $\hat{\theta}$. The resulting bootstrap method is known as the *percentile t-method*, in view of the fact that it is based on estimating quantiles of the "studentized" statistic $(\hat{\theta} - \theta)/\hat{\sigma}$. (The notion of a t-statistic is used here in an abstract manner to denote a centered statistic divided by a scale estimate; in general, there is no relationship with Student's t-distribution from normal theory.) A simpler method is to choose $\hat{\sigma}$ independent of the data. If we choose $\hat{\sigma} = \hat{\sigma}^* = 1$, then the bootstrap quantiles $\hat{\xi}_\alpha$ are the quantiles of the centered statistic $\hat{\theta}^* - \hat{\theta}$. This is known as the *percentile method*. Both methods yield asymptotically correct confidence levels, although the percentile t-method is generally more accurate.

A third method, *Efron's percentile method*, proposes the confidence interval $[\hat{\zeta}_{1-\beta}, \hat{\zeta}_\alpha]$ for $\hat{\zeta}_\alpha$ equal to the upper α-quantile of $\hat{\theta}^*$: the smallest value $x = \hat{\zeta}_\alpha$ such that

$$\mathrm{P}(\hat{\theta}^* \le x \mid \hat{P}) \ge 1 - \alpha.$$

Thus, $\hat{\zeta}_\alpha$ results from "bootstrapping" $\hat{\theta}$, while $\hat{\xi}_\alpha$ is the product of bootstrapping $(\hat{\theta} - \theta)/\hat{\sigma}$. These quantiles are related, and Efron's percentile interval can be reexpressed in the quantiles $\hat{\xi}_\alpha$ of $\hat{\theta}^* - \hat{\theta}$ (employed by the percentile method with $\hat{\sigma} = 1$) as

$$[\hat{\zeta}_{1-\beta}, \hat{\zeta}_\alpha] = [\hat{\theta} + \hat{\xi}_{1-\beta}, \hat{\theta} + \hat{\xi}_\alpha].$$

The logical justification for this interval is less strong than for the intervals based on bootstrapping $\hat{\theta} - \theta$, but it appears to work well. The two types of intervals coincide in the case that the conditional distribution of $\hat{\theta}^* - \hat{\theta}$ is symmetric about zero. We shall see that the difference is asymptotically negligible if $\hat{\theta}^* - \hat{\theta}$ converges to a normal distribution.

Efron's percentile interval is the only one among the three intervals that is invariant under monotone transformations. For instance, if setting a confidence interval for the correlation coefficient, the sample correlation coefficient might be transformed by Fisher's transformation before carrying out the bootstrap scheme. Next, the confidence interval for the transformed correlation can be transformed back into a confidence interval for the correlation coefficient. This operation would have no effect on Efron's percentile interval but can improve the other intervals considerably, in view of the skewness of the statistic. In this sense Efron's method automatically "finds" useful (stabilizing) transformations. The fact that it does not become better through transformations of course does not imply that it is good, but the invariance appears desirable.

Several of the elements of the bootstrap scheme are still unspecified. The missing probability $\alpha + \beta$ can be distributed over the two tails of the confidence interval in several ways. In many situations equal-tailed confidence intervals, corresponding to the choice $\alpha = \beta$, are reasonable. In general, these do not have $\hat{\theta}$ exactly as the midpoint of the interval. An alternative is the interval

$$\left[\hat{\theta} - \hat{\xi}^s_{\alpha+\beta}\hat{\sigma}, \hat{\theta} + \hat{\xi}^s_{\alpha+\beta}\hat{\sigma}\right],$$

with $\hat{\xi}^s_\alpha$ equal to the upper α-quantile of $|\hat{\theta}^* - \hat{\theta}|/\hat{\sigma}^*$. A further possibility is to choose α and β under the side condition that the difference $\hat{\xi}_\beta - \hat{\xi}_{1-\alpha}$, which is proportional to the length of the confidence interval, is minimal.

More interesting is the choice of the estimator $\hat{P}$ for the underlying distribution. If the original observations are a random sample $X_1, \ldots, X_n$ from a probability distribution P, then one candidate is the empirical distribution $\mathbb{P}_n = n^{-1}\sum \delta_{X_i}$ of the observations, leading to the *empirical bootstrap*. Generating a random sample from the empirical distribution amounts to resampling with replacement from the set $\{X_1, \ldots, X_n\}$ of original observations. The name "bootstrap" derives from this resampling procedure, which might be surprising at first, because the observations are "sampled twice." If we view the bootstrap as a nonparametric plug-in estimator, we see that there is nothing peculiar about resampling.

We shall be mostly concerned with the empirical bootstrap, even though there are many other possibilities. If the observations are thought to follow a specified parametric model, then it is more reasonable to set $\hat{P}$ equal to $P_{\hat{\theta}}$ for a given estimator $\hat{\theta}$. This is what one would have done in the first place, but it is called the *parametric bootstrap* within the present context. That the bootstrapping methodology is far from obvious is clear from the fact that the literature also considers the exchangeable, the Bayesian, the smoothed, and the wild bootstrap, as well as several schemes for bootstrap corrections. Even "resampling" can be carried out differently, for instance, by sampling fewer than n variables, or without replacement.

It is almost never possible to calculate the bootstrap quantiles $\hat{\xi}_\alpha$ numerically. In practice, these estimators are approximated by a simulation procedure. A large number of independent bootstrap samples $X_1^*, \ldots, X_n^*$ are generated according to the estimated distribution $\hat{P}$. Each sample gives rise to a bootstrap value $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ of the standardized statistic. Finally, the bootstrap quantiles $\hat{\xi}_\alpha$ are estimated by the empirical quantiles of these bootstrap

values. This simulation scheme always produces an additional (random) error in the coverage probability of the resulting confidence interval. In principle, by using a sufficiently large number of bootstrap samples, possibly combined with an efficient method of simulation, this error can be made arbitrarily small. Therefore the additional error is usually ignored in the theory of the bootstrap procedure. This chapter follows this custom and concerns the "exact" distribution and quantiles of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, without taking a simulation error into account.

23.2 Consistency

A confidence interval $[\hat{\theta}_{n,1}, \hat{\theta}_{n,2}]$ is (conservatively) *asymptotically consistent* at level $1 - \alpha - \beta$ if, for every possible P,

$$\liminf_{n\to\infty} \mathrm{P}(\hat{\theta}_{n,1} \le \theta \le \hat{\theta}_{n,2} \mid P) \ge 1 - \alpha - \beta.$$

The consistency of a bootstrap confidence interval is closely related to the consistency of the bootstrap estimator of the distribution of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$. The latter is best defined relative to a metric on the collection of possible laws of the estimator. Call the bootstrap estimator for the distribution *consistent* relative to the Kolmogorov-Smirnov distance if

$$\sup_x \left| \mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le x \mid P \right) - \mathrm{P}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*} \le x \mid \hat{P}_n \right) \right| \xrightarrow{\mathrm{P}} 0.$$

It is not a great loss of generality to assume that the sequence $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ converges in distribution to a continuous distribution function F (in our examples Φ). Then consistency relative to the Kolmogorov-Smirnov distance is equivalent to the requirements, for every x,

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le x \mid P \right) \to F(x), \qquad \mathrm{P}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*} \le x \mid \hat{P}_n \right) \xrightarrow{\mathrm{P}} F(x). \tag{23.2}$$

(See Problem 23.1.) This type of consistency implies the asymptotic consistency of confidence intervals.

23.3 ***Lemma.*** *Suppose that $(\hat{\theta}_n - \theta)/\hat{\sigma}_n \rightsquigarrow T$, and that $(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^* \rightsquigarrow T$ given the original observations, in probability, for a random variable T with a continuous distribution function. Then the bootstrap confidence intervals $[\hat{\theta}_n - \hat{\xi}_{n,\beta}\hat{\sigma}_n, \hat{\theta}_n - \hat{\xi}_{n,1-\alpha}\hat{\sigma}_n]$ are asymptotically consistent at level $1 - \alpha - \beta$. If the conditions hold for nonrandom $\hat{\sigma}_n = \hat{\sigma}_n^*$, and T is symmetrically distributed about zero, then the same is true for Efron's percentile intervals.*

Proof. Every subsequence has a further subsequence along which the sequence $(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ converges weakly to T, conditionally, almost surely. For simplicity, assume that the whole sequence converges almost surely; otherwise, argue along subsequences.

If a sequence of distribution functions F_n converges weakly to a distribution function F, then the corresponding quantile functions F_n^{-1} converge to the quantile function F^{-1} at every continuity point (see Lemma 21.2). Apply this to the (random) distribution functions $\hat{F}_n$ of $(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ and a continuity point $1 - \alpha$ of the quantile function F^{-1} of T to conclude

that $\hat{\xi}_{n,\alpha} = \hat{F}_n^{-1}(1-\alpha)$ converges almost surely to $F^{-1}(1-\alpha)$. By Slutsky's lemma, the sequence $(\hat{\theta}_n - \theta)/\hat{\sigma}_n - \hat{\xi}_{n,\alpha}$ converges weakly to $T - F^{-1}(1-\alpha)$. Thus

$$\mathrm{P}(\theta \geq \hat{\theta}_n - \hat{\sigma}_n \hat{\xi}_{n,\alpha}) = \mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq \hat{\xi}_{n,\alpha} \mid P\right) \to \mathrm{P}\big(T \leq F^{-1}(1-\alpha)\big) = 1 - \alpha.$$

This argument applies to all except at most countably many α. Because both the left and the right sides of the preceding display are monotone functions of α and the right side is continuous, it must be valid for every α. The consistency of the bootstrap confidence interval follows.

Efron's percentile interval is the interval $[\hat{\zeta}_{n,1-\beta}, \hat{\zeta}_{n,\alpha}]$, where $\hat{\zeta}_{n,\alpha} = \hat{\theta}_n + \hat{\xi}_{n,\alpha}]$. By the preceding argument,

$$\mathrm{P}(\theta \geq \hat{\zeta}_{n,1-\beta}) = \mathrm{P}(\hat{\theta}_n - \theta \leq -\hat{\xi}_{n,1-\beta} \mid P) \to \mathrm{P}\big(T \leq -F^{-1}(\beta)\big) = 1 - \beta.$$

The last equality follows by the symmetry of T. The consistency follows. ■

From now on we consider the empirical bootstrap; that is, $\hat{P}_n = \mathbb{P}_n$ is the empirical distribution of a random sample $X_1, \ldots, X_n$. We shall establish (23.2) for a large class of statistics, with F the normal distribution. Our method is first to prove the consistency for $\hat{\theta}_n$ equal to the sample mean and next to show that the consistency is retained under application of the delta method. Combining these results, we obtain the consistency of many bootstrap procedures, for instance for setting confidence intervals for the correlation coefficient.

In view of Slutsky's lemma, weak convergence of the centered sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ combined with convergence in probability of $\hat{\sigma}_n/\sqrt{n}$ yields the weak convergence of the studentized statistics $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$. An analogous statement is true for the bootstrap statistic, for which the convergence in probability of $\hat{\sigma}_n^*/\sqrt{n}$ must be shown conditionally on the original observations. Establishing (conditional) consistency of $\hat{\sigma}_n/\sqrt{n}$ and $\hat{\sigma}_n^*/\sqrt{n}$ is usually not hard. Therefore, we restrict ourselves to studying the nonstudentized statistics.

Let $\overline{X}_n$ be the mean of a sample of n random vectors from a distribution with finite mean vector μ and covariance matrix Σ. According to the multivariate central limit theorem, the sequence $\sqrt{n}(\overline{X}_n - \mu)$ is asymptotically normal $N(0, \Sigma)$-distributed. We wish to show the same for $\sqrt{n}(\overline{X}_n^* - \overline{X}_n)$, in which $\overline{X}_n^*$ is the average of n observations from $\mathbb{P}_n$, that is, of n values resampled from the set of original observations $\{X_1, \ldots, X_n\}$ with replacement.

23.4 ***Theorem (Sample mean).*** *Let $X_1, X_2, \ldots$ be i.i.d. random vectors with mean μ and covariance matrix Σ. Then conditionally on $X_1, X_2, \ldots$, for almost every sequence $X_1, X_2, \ldots$,*

$$\sqrt{n}(\overline{X}_n^* - \overline{X}_n) \rightsquigarrow N(0, \Sigma).$$

Proof. For a fixed sequence $X_1, X_2, \ldots$, the variable $\overline{X}_n^*$ is the average of n observations $X_1^*, \ldots, X_n^*$ sampled from the empirical distribution $\mathbb{P}_n$. The (conditional) mean and covariance matrix of these observations are

$$\mathrm{E}(X_i^* \mid \mathbb{P}_n) = \sum_{i=1}^n \frac{1}{n} X_i = \overline{X}_n,$$

$$\mathrm{E}\big((X_i^* - \overline{X}_n)(X_i^* - \overline{X}_n)^T \mid \mathbb{P}_n\big) = \sum_{i=1}^n \frac{1}{n}(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T$$

$$= \overline{X_n X_n^T} - \overline{X}_n \overline{X}_n^T.$$

By the strong law of large numbers, the conditional covariance converges to Σ for almost every sequence $X_1, X_2, \ldots$.

The asymptotic distribution of $\overline{X}_n^*$ can be established by the central limit theorem. Because the observations $X_1^*, \ldots, X_n^*$ are sampled from a different distribution $\mathbb{P}_n$ for every n, a central limit theorem for a triangular array is necessary. The Lindeberg central limit theorem, Theorem 2.27, is appropriate. It suffices to show that, for every $\varepsilon > 0$,

$$\mathrm{E}\|X_i^*\|^2 1\{\|X_i^*\| > \varepsilon\sqrt{n}\} = \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 1\{\|X_i\| > \varepsilon\sqrt{n}\} \xrightarrow{\text{as}} 0.$$

The left side is smaller than $n^{-1}\sum_{i=1}^n \|X_i\|^2 1\{\|X_i\| > M\}$ as soon as $\varepsilon\sqrt{n} \geq M$. By the strong law of large numbers, the latter average converges to $\mathrm{E}\|X_i\|^2 1\{\|X_i\| > M\}$ for almost every sequence $X_1, X_2, \ldots$. For sufficiently large M, this expression is arbitrarily small. Conclude that the limit superior of the left side of the preceding display is smaller than any number $\eta > 0$ almost surely and hence the left side converges to zero for almost every sequence $X_1, X_2, \ldots$. ■

Assume that $\hat{\theta}_n$ is a statistic, and that ϕ is a given differentiable map. If the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution, then so does the sequence $\sqrt{n}\big(\phi(\hat{\theta}_n) - \phi(\theta)\big)$, by the delta method. The bootstrap estimator for the distribution of $\phi(\hat{\theta}_n) - \phi(\theta)$ is $\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)$. If the bootstrap is consistent for estimating the distribution of $\hat{\theta}_n - \theta$, then it is also consistent for estimating the distribution of $\phi(\hat{\theta}_n) - \phi(\theta)$.

23.5 ***Theorem (Delta method for bootstrap).*** *Let $\phi: \mathbb{R}^k \mapsto \mathbb{R}^m$ be a measurable map defined and continuously differentiable in a neighborhood of θ. Let $\hat{\theta}_n$ be random vectors taking their values in the domain of ϕ that converge almost surely to θ. If $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow T$, and $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \rightsquigarrow T$ conditionally almost surely, then both $\sqrt{n}\big(\phi(\hat{\theta}_n) - \phi(\theta)\big) \rightsquigarrow \phi_\theta'(T)$ and $\sqrt{n}\big(\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\big) \rightsquigarrow \phi_\theta'(T)$, conditionally almost surely.*

Proof. By the mean value theorem, the difference $\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)$ can be written as $\phi_{\tilde{\theta}_n}'(\hat{\theta}_n^* - \hat{\theta}_n)$ for a point $\tilde{\theta}_n$ between $\hat{\theta}_n^*$ and $\hat{\theta}_n$, if the latter two points are in the ball around θ in which ϕ is continuously differentiable. By the continuity of the derivative, there exists for every $\eta > 0$ a constant $\delta > 0$ such that $\|\phi_{\theta'}'h - \phi_\theta'h\| < \eta\|h\|$ for every h and every $\|\theta' - \theta\| \leq \delta$. If n is sufficiently large, δ sufficiently small, $\sqrt{n}\|\hat{\theta}_n^* - \hat{\theta}_n\| \leq M$, and $\|\hat{\theta}_n - \theta\| \leq \delta$, then

$$\begin{aligned} R_n :&= \left\|\sqrt{n}\big(\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\big) - \phi_\theta'\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\right\| \\ &= \left|(\phi_{\tilde{\theta}_n}' - \phi_\theta')\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\right| \leq \eta M. \end{aligned}$$

Fix a number $\varepsilon > 0$ and a large number M. For η sufficiently small to ensure that $\eta M < \varepsilon$,

$$\mathrm{P}(R_n > \varepsilon \mid \hat{P}_n) \leq \mathrm{P}\big(\sqrt{n}\|\hat{\theta}_n^* - \hat{\theta}_n\| > M \text{ or } \|\hat{\theta}_n - \theta\| > \delta \mid \hat{P}_n\big).$$

Because $\hat{\theta}_n$ converges almost surely to θ, the right side converges almost surely to $\mathrm{P}\big(\|T\| \geq M\big)$ for every continuity point M of $\|T\|$. This can be made arbitrarily small by choice of M. Conclude that the left side converges to 0 almost surely. The theorem follows by an application of Slutsky's lemma. ■

23.6 ***Example (Sample variance).*** The (biased) sample variance $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ equals $\phi(\overline{X}_n, \overline{X_n^2})$ for the map $\phi(x, y) = y - x^2$. The empirical bootstrap is consistent

for estimation of the distribution of $(\overline{X}_n, \overline{X_n^2}) - (\alpha_1, \alpha_2)$, by Theorem 23.4, provided that the fourth moment of the underlying distribution is finite. The delta method shows that the empirical bootstrap is consistent for estimating the distribution of $S_n^2 - \sigma^2$ in that

$$\sup_x \left| \mathrm{P}\Big(\sqrt{n}\big(S_n^2 - \sigma^2\big) \le x \mid P\Big) - \mathrm{P}\Big(\sqrt{n}\big(S_n^{*2} - S_n^2\big) \le x \mid \mathbb{P}_n\Big) \right| \stackrel{\text{as}}{\to} 0.$$

The asymptotic variance of S_n^2 can be estimated by $S_n^4(k_n + 2)$, in which k_n is the sample kurtosis. The law of large numbers shows that this estimator is asymptotically consistent. The bootstrap version of this estimator can be shown to be consistent given almost every sequence of the original observations. Thus, the consistency of the empirical bootstrap extends to the studentized statistic $(S_n^2 - \sigma^2)/S_n^2\sqrt{k_n + 1}$. □

*23.2.1 *Empirical Bootstrap*

In this section we follow the same method as previously, but we replace the sample mean by the empirical distribution and the delta method by the functional delta method. This is more involved, but more flexible, and yields, for instance, the consistency of the bootstrap of the sample median.

Let $\mathbb{P}_n$ be the empirical distribution of a random sample $X_1, \ldots, X_n$ from a distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a Donsker class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, as defined in Chapter 19. Given the sample values, let $X_1^*, \ldots, X_n^*$ be a random sample from $\mathbb{P}_n$. The *bootstrap empirical distribution* is the empirical measure $\mathbb{P}_n^* = n^{-1}\sum_{i=1}^n \delta_{X_i^*}$, and the *bootstrap empirical process* $\mathbb{G}_n^*$ is

$$\mathbb{G}_n^* = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1)\, \delta_{X_i},$$

in which M_{ni} is the number of times that X_i is "redrawn" from $\{X_1, \ldots, X_n\}$ to form $X_1^*, \ldots, X_n^*$. By construction, the vector of counts $(M_{n1}, \ldots, M_{nn})$ is independent of $X_1, \ldots, X_n$ and multinomially distributed with parameters n and (probabilities) $1/n, \ldots, 1/n$.

If the class $\mathcal{F}$ has a finite envelope function F, then both the empirical process $\mathbb{G}_n$ and the bootstrap process $\mathbb{G}_n^*$ can be viewed as maps into the space $\ell^\infty(\mathcal{F})$. The analogue of Theorem 23.4 is that the sequence $\mathbb{G}_n^*$ converges in $\ell^\infty(\mathcal{F})$ conditionally in distribution to the same limit as the sequence $\mathbb{G}_n$, a tight Brownian bridge process $\mathbb{G}_P$. To give a precise meaning to "conditional weak convergence" in $\ell^\infty(\mathcal{F})$, we use the *bounded Lipschitz metric*. It can be shown that a sequence of random elements in $\ell^\infty(\mathcal{F})$ converges in distribution to a tight limit in $\ell^\infty(\mathcal{F})$ if and only if†

$$\sup_{h \in \mathrm{BL}_1(\ell^\infty(\mathcal{F}))} \left| \mathrm{E}^* h(G_n) - \mathrm{E} h(G) \right| \to 0.$$

We use the notation E_M to denote "taking the expectation conditionally on $X_1, \ldots, X_n$," or the expectation with respect to the multinomial vectors M_n only.‡

† For a metric space $\mathbb{D}$, the set $\mathrm{BL}_1(\mathbb{D})$ consists of all functions $h : \mathbb{D} \mapsto [-1, 1]$ that are uniformly Lipschitz: $|h(z_1) - h(z_2)| \le d(z_1, z_2)$ for every pair (z_1, z_2). See, for example, Chapter 1.12 of [146].

‡ For a proof of Theorem 23.7, see the original paper [58], or, for example, Chapter 3.6 of [146].

23.7 *Theorem (Empirical bootstrap).* *For every Donsker class $\mathcal{F}$ of measurable functions with finite envelope function F,*

$$\sup_{h\in \mathrm{BL}_1(\ell^\infty(\mathcal{F}))} \left|\mathrm{E}_M h(\mathbb{G}_n^*) - \mathrm{E}h(\mathbb{G}_P)\right| \xrightarrow{\mathrm{P}} 0.$$

Furthermore, the sequence $\mathbb{G}_n^*$ is asymptotically measurable. If $P^*F^2 < \infty$, then the convergence is outer almost surely as well.

Next, consider an analogue of Theorem 23.5, using the functional delta method. Theorem 23.5 goes through without too many changes. However, for many infinite-dimensional applications of the delta method the condition of *continuous* differentiability imposed in Theorem 23.5 fails. This problem may be overcome in several ways. In particular, continuous differentiability is not necessary for the consistency of the bootstrap "in probability" (rather than "almost surely"). Because this appears to be sufficient for statistical applications, we shall limit ourselves to this case.

Consider sequences of maps $\hat\theta_n$ and $\hat\theta_n^*$ with values in a normed space $\mathbb{D}$ (e.g., $\ell^\infty(\mathcal{F})$) such that the sequence $\sqrt{n}(\hat\theta_n - \theta)$ converges unconditionally in distribution to a tight random element T, and the sequence $\sqrt{n}(\hat\theta_n^* - \hat\theta_n)$ converges conditionally given $X_1, X_2, \ldots$ in distribution to the same random element T. A precise formulation of the second is that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}_M h\big(\sqrt{n}(\hat\theta_n^* - \hat\theta_n)\big) - \mathrm{E}h(T)\right| \xrightarrow{\mathrm{P}} 0. \tag{23.8}$$

Here the notation E_M means the conditional expectation given the original data $X_1, X_2, \ldots$ and is motivated by the application to the bootstrap empirical distribution.† By the preceding theorem, the empirical distribution $\hat\theta_n = \mathbb{P}_n$ satisfies condition (23.8) if viewed as a map in $\ell^\infty(\mathcal{F})$ for a Donsker class $\mathcal{F}$.

23.9 *Theorem (Delta method for bootstrap).* *Let $\mathbb{D}$ be a normed space and let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{R}^k$ be Hadamard differentiable at θ tangentially to a subspace $\mathbb{D}_0$. Let $\hat\theta_n$ and $\hat\theta_n^*$ be maps with values in $\mathbb{D}_\phi$ such that $\sqrt{n}(\hat\theta_n - \theta) \rightsquigarrow T$ and such that (23.8) holds, in which $\sqrt{n}(\hat\theta_n^* - \hat\theta_n)$ is asymptotically measurable and T is tight and takes its values in $\mathbb{D}_0$. Then the sequence $\sqrt{n}\big(\phi(\hat\theta_n^*) - \phi(\hat\theta_n)\big)$ converges conditionally in distribution to $\phi'_\theta(T)$, given $X_1, X_2, \ldots$, in probability.*

Proof. By the Hahn-Banach theorem it is not a loss of generality to assume that the derivative $\phi'_\theta : \mathbb{D} \mapsto \mathbb{R}^k$ is defined and continuous on the whole space. For every $h \in \mathrm{BL}_1(\mathbb{R}^k)$, the function $h \circ \phi'_\theta$ is contained in $\mathrm{BL}_{\|\phi'_\theta\|}(\mathbb{D})$. Thus (23.8) implies

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^k)} \left|\mathrm{E}_M h\Big(\phi'_\theta\big(\sqrt{n}(\hat\theta_n^* - \hat\theta_n)\big)\Big) - \mathrm{E}h\big(\phi'_\theta(T)\big)\right| \xrightarrow{\mathrm{P}} 0.$$

Because $\left|h(x) - h(y)\right|$ is bounded by $2 \wedge d(x, y)$ for every $h \in \mathrm{BL}_1(\mathbb{R}^k)$,

$$\begin{aligned}\sup_{h\in \mathrm{BL}_1(\mathbb{R}^k)} &\left|\mathrm{E}_M h\Big(\sqrt{n}\big(\phi(\hat\theta_n^*) - \phi(\hat\theta_n)\big)\Big) - \mathrm{E}_M h\Big(\phi'_\theta\big(\sqrt{n}(\hat\theta_n^* - \hat\theta_n)\big)\Big)\right| \\ &\le \varepsilon + 2\mathrm{P}_M\left(\left\|\sqrt{n}\big(\phi(\hat\theta_n^*) - \phi(\hat\theta_n)\big) - \phi'_\theta\big(\sqrt{n}(\hat\theta_n^* - \hat\theta_n)\big)\right\| > \varepsilon\right).\end{aligned} \tag{23.10}$$

† It is assumed that $h\big(\sqrt{n}(\hat\theta_n^* - \hat\theta_n)\big)$ is a measurable function of M.

The theorem is proved once it has been shown that the conditional probability on the right converges to zero in outer probability.

The sequence $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n, \hat{\theta}_n - \theta)$ converges (unconditionally) in distribution to a pair of two independent copies of T. This follows, because conditionally given $X_1, X_2, \ldots$, the second component is deterministic, and the first component converges in distribution to T, which is the same for every sequence $X_1, X_2, \ldots$. Therefore, by the continuous-mapping theorem both sequences $\sqrt{n}(\hat{\theta}_n - \theta)$ and $\sqrt{n}(\hat{\theta}_n^* - \theta)$ converge (unconditionally) in distribution to separable random elements that concentrate on the linear space $\mathbb{D}_0$. By Theorem 20.8,

$$\sqrt{n}\big(\phi(\hat{\theta}_n^*) - \phi(\theta)\big) = \phi_\theta'\big(\sqrt{n}(\hat{\theta}_n^* - \theta)\big) + o_P^*(1),$$
$$\sqrt{n}\big(\phi(\hat{\theta}_n) - \phi(\theta)\big) = \phi_\theta'\big(\sqrt{n}(\hat{\theta}_n - \theta)\big) + o_P^*(1).$$

Subtract the second from the first equation to conclude that the sequence $\sqrt{n}\big(\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\big) - \phi_\theta'\big(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)\big)$ converges (unconditionally) to 0 in outer probability. Thus, the conditional probability on the right in (23.10) converges to zero in outer mean. This concludes the proof. ■

23.11 ***Example (Empirical distribution function).*** Because the cells $(-\infty, t] \subset \mathbb{R}$ form a Donsker class, the empirical distribution function $\mathbb{F}_n$ of a random sample of real-valued variables satisfies the condition of the preceding theorem. Thus, conditionally on $X_1, X_2, \ldots$, the sequence $\sqrt{n}\big(\phi(\mathbb{F}_n^*) - \phi(\mathbb{F}_n)\big)$ converges in distribution to the same limit as $\sqrt{n}\big(\phi(\mathbb{F}_n) - \phi(F)\big)$, for every Hadamard-differentiable function ϕ.

This includes, among others, quantiles and trimmed means, under the same conditions on the underlying measure F that ensure that empirical quantiles and trimmed means are asymptotically normal. See Lemmas 21.3, 22.9, and 22.10. □

23.3 Higher-Order Correctness

The investigation of the performance of a bootstrap confidence interval can be refined by taking into account the order at which the true level converges to the desired level. A confidence interval is (conservatively) *correct at level* $1 - \alpha - \beta$ *up to order* $O(n^{-k})$ if

$$\mathrm{P}\big(\hat{\theta}_{n,1} \le \theta \le \hat{\theta}_{n,2} \mid P\big) \ge 1 - \alpha - \beta - O\left(\frac{1}{n^k}\right).$$

Similarly, the quality of the bootstrap estimator for distributions can be assessed more precisely by the rate at which the Kolmogorov-Smirnov distance between the distribution function of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ and the conditional distribution function of $(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ converges to zero. We shall see that the percentile t-method usually performs better than the percentile method. For the percentile t-method, the Kolmogorov-Smirnov distance typically converges to zero at the rate $O_P(n^{-1})$, whereas the percentile method attains "only" an $O_P(n^{-1/2})$ rate of correctness. The latter is comparable to the error of the normal approximation.

Rates for the Kolmogorov-Smirnov distance translate directly into orders of correctness of one-tailed confidence intervals. The correctness of two-tailed or symmetric confidence intervals may be higher, because of the cancellation of the coverage errors contributed by

the left and right tails. In many cases the percentile method, the percentile t-method, and the normal approximation all yield correct two-tailed confidence intervals up to order $O(n^{-1})$. Their relative qualities may be studied by a more refined analysis. This must also take into account the length of the confidence intervals, for an increase in length of order $O_P(n^{-3/2})$ may easily reduce the coverage error to the order $O(n^{-k})$ for any k.

The technical tool to obtain these results are *Edgeworth expansions*. Edgeworth's classical expansion is a refinement of the central limit theorem that shows the magnitude of the difference between the distribution function of a sample mean and its normal approximation. Edgeworth expansions have subsequently been obtained for many other statistics as well.

An Edgeworth expansion for the distribution function of a statistic $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ is typically an expansion in increasing powers of $1/\sqrt{n}$ of the form

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le x \mid P\right) = \Phi(x) + \frac{p_1(x \mid P)}{\sqrt{n}}\phi(x) + \frac{p_2(x \mid P)}{n}\phi(x) + \cdots \tag{23.12}$$

The remainder is of lower order than the last included term, uniformly in the argument x. Thus, in the present case the remainder is $o(n^{-1})$ (or even $O(n^{-3/2})$). The functions p_i are polynomials in x, whose coefficients depend on the underlying distribution, typically through (asymptotic) moments of the pair $(\hat{\theta}_n, \hat{\sigma}_n)$.

23.13 ***Example (Sample mean).*** Let $\overline{X}_n$ be the mean of a random sample of size n, and let $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ be the (biased) sample variance. If μ, σ^2, λ and κ are the mean, variance, skewness and kurtosis of the underlying distribution, then

$$\mathrm{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le x \mid P\right) = \Phi(x) - \frac{\lambda(x^2-1)}{6\sqrt{n}}\phi(x) - \frac{3\kappa(x^3-3x) + \lambda^2(x^5 - 10x^3 + 15x)}{72n}\phi(x) + O\left(\frac{1}{n\sqrt{n}}\right).$$

These are the first two terms of the classical expansion of Edgeworth. If the standard deviation of the observations is unknown, an Edgeworth expansion of the t-statistic is of more interest. This takes the form (see [72, pp. 71–73])

$$\mathrm{P}\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \le x \mid P\right) = \Phi(x) + \frac{\lambda(2x^2+1)}{6\sqrt{n}}\phi(x) + \frac{3\kappa(x^3-3x) - 2\lambda^2(x^5 + 2x^3 - 3x) - 9(x^3+3x)}{36n}\phi(x) + O\left(\frac{1}{n\sqrt{n}}\right).$$

Although the polynomials are different, these expansions are of the same form. Note that the polynomial appearing in the $1/\sqrt{n}$ term is even in both cases.

These expansions generally fail if the underlying distribution of the observations is discrete. *Cramér's condition* requires that the modulus of the characteristic function of the observations be bounded away from unity on closed intervals that do not contain the origin. This condition is satisfied if the observations possess a density with respect to Lebesgue measure. Next to Cramér's condition a sufficient number of moments of the observations must exist. □

23.14 ***Example (Studentized quantiles).*** The pth quantile $F^{-1}(p)$ of a distribution function F may be estimated by the empirical pth quantile $\mathbb{F}_n^{-1}(p)$. This is the rth order statistic of the sample for r equal to the largest integer not greater than np. Its mean square error can be computed as

$$\mathrm{E}\big(\mathbb{F}_n^{-1}(p) - F^{-1}(p)\big)^2 = r\binom{n}{r}\int_0^1 \big(F^{-1}(u) - F^{-1}(p)\big)^2 u^{r-1}(1-u)^{n-r}\,du.$$

An empirical estimator $\hat{\sigma}_n$ for the mean square error of $\mathbb{F}_n^{-1}(p)$ is obtained by replacing F by the empirical distribution function. If the distribution has a differentiable density f, then

$$\mathrm{P}\left(\frac{\mathbb{F}_n^{-1}(p) - F^{-1}(p)}{\hat{\sigma}_n} \le x \mid F\right) = \Phi(x) + \frac{p_1(x \mid F)}{\sqrt{n}}\phi(x) + O\left(\frac{1}{n^{3/4}}\right),$$

where $p_1(x \mid F)$ is the polynomial of degree 3 given by (see [72, pp. 318–321])

$$p_1(x \mid F)\, 12\sqrt{p(1-p)} = \frac{3}{\sqrt{\pi}}x^3 + \left[2 - 10p - 12p(1-p)\frac{f'}{f^2}\big(F^{-1}(p)\big)\right]x^2 + \frac{3+6\sqrt{2}}{\sqrt{\pi}}x - 8 + 4p - 12(r - np).$$

This expansion is unusual in two respects. First, the remainder is of the order $O(n^{-3/4})$ rather than of the order $O(n^{-1})$. Second, the polynomial appearing in the first term is not even. For this reason several of the conclusions of this section are not valid for sample quantiles. In particular, the order of correctness of all empirical bootstrap procedures is $O_P(n^{-1/2})$, not greater. In this case, a "smoothed bootstrap" based on "resampling" from a density estimator (as in Chapter 24) may be preferable, depending on the underlying distribution. □

If the distribution function of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ admits an Edgeworth expansion (23.12), then it is immediate that the normal approximation is correct up to order $O(1/\sqrt{n})$. Evaluation of the expansion at the normal quantiles z_β and $z_{1-\alpha}$ yields

$$\mathrm{P}(\hat{\theta}_n - z_\beta\,\hat{\sigma}_n \le \theta \le \hat{\theta}_n - z_{1-\alpha}\,\hat{\sigma}_n \mid P) = 1 - \alpha - \beta + \frac{p_1(z_\beta \mid P)\phi(z_\beta) - p_1(z_{1-\alpha} \mid P)\phi(z_{1-\alpha})}{\sqrt{n}} + O\left(\frac{1}{n}\right).$$

Thus, the level of the confidence interval $[\hat{\theta}_n - z_\beta\,\hat{\sigma}_n, \hat{\theta}_n - z_{1-\alpha}\,\hat{\sigma}_n]$ is $1 - \alpha - \beta$ up to order $O(1/\sqrt{n})$. For a two-tailed, symmetric interval, α and β are chosen equal. Inserting $z_\beta = z_\alpha = -z_{1-\alpha}$ in the preceding display, we see that the errors of order $1/\sqrt{n}$ resulting from the left and right tails cancel each other if p_1 is an even function. In this common situation the order of correctness improves to $O(n^{-1})$.

It is of theoretical interest that the coverage probability can be corrected up to any order by making the normal confidence interval slightly wider than first-order asymptotics would suggest. The interval may be widened by using quantiles z_{α_n} with $\alpha_n < \alpha$, rather than z_α. In view of the preceding display, for any α_n,

$$\mathrm{P}(\hat{\theta}_n - z_{\alpha_n}\,\hat{\sigma}_n \le \theta \le \hat{\theta}_n - z_{1-\alpha_n}\,\hat{\sigma}_n \mid P) = 1 - 2\alpha_n + O\left(\frac{1}{n}\right).$$

The $O(n^{-1})$ term results from the Edgeworth expansion (23.12) and is universal, independent of the sequence α_n. For $\alpha_n = \alpha - M/n$ and a sufficiently large constant M, the right side becomes

$$1 - 2\alpha + \frac{2M}{n} + O\left(\frac{1}{n}\right) \geq 1 - 2\alpha - O\left(\frac{1}{n^k}\right).$$

Thus, a slight widening of the normal confidence interval yields asymptotically correct (conservative) coverage probabilities up to any order $O(n^{-k})$. If $\hat{\sigma}_n = O_P(n^{-1/2})$, then the widened interval is $2(z_{\alpha_n} - z_\alpha)\,\hat{\sigma}_n = O_P(n^{-3/2})$ wider than the normal confidence interval. This difference is small relatively to the absolute length of the interval, which is $O_P(n^{-1/2})$. Also, the choice of the scale estimator $\hat{\sigma}_n$ (which depends on $\hat{\theta}_n$) influences the width of the interval stronger than replacing ξ_α by ξ_{α_n}.

An Edgeworth expansion usually remains valid in a conditional sense if a good estimator $\hat{P}_n$ is substituted for the true underlying distribution P. The bootstrap version of expansion (23.12) is

$$\mathrm{P}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*} \leq x \mid \hat{P}_n\right) = \Phi(x) + \frac{p_1(x \mid \hat{P}_n)}{\sqrt{n}}\phi(x) + \frac{p_2(x \mid \hat{P}_n)}{n}\phi(x) + \cdots.$$

In this expansion the remainder term is a random variable, which ought to be of smaller order in probability than the last term. In the given expansion the remainder ought to be $o_P(n^{-1})$ uniformly in x. Subtract the bootstrap expansion from the unconditional expansion (23.12) to obtain that

$$\begin{aligned}&\sup_x \left|\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq x \mid P\right) - \mathrm{P}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*} \leq x \mid \hat{P}_n\right)\right| \\ &\qquad \leq \sup_x \left|\frac{p_1(x \mid P) - p_1(x \mid \hat{P}_n)}{\sqrt{n}} + \frac{p_2(x \mid P) - p_2(x \mid \hat{P}_n)}{n}\right| \phi(x) + o_P\left(\frac{1}{n}\right).\end{aligned}$$

The polynomials p_i typically depend on P in a smooth way, and the difference $\hat{P}_n - \hat{P}$ is typically of the order $O_P(n^{-1/2})$. Then the Kolmogorov-Smirnov distance between the true distribution function of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ and its percentile t-bootstrap estimator is of the order $O_P(n^{-1})$.

The analysis of the percentile method starts from an Edgeworth expansion of the distribution function of the unstudentized statistic $\hat{\theta}_n - \theta$. This has as leading term the normal distribution with variance σ_n^2, the asymptotic variance of $\hat{\theta}_n - \theta$, rather than the standard normal distribution. Typically it is of the form

$$\begin{aligned}\mathrm{P}(\hat{\theta}_n - \theta \leq x \mid P) = \Phi\left(\frac{x}{\sigma_n}\right) &+ \frac{1}{\sqrt{n}} q_1\left(\frac{x}{\sigma_n} \,\middle|\, P\right)\phi\left(\frac{x}{\sigma_n}\right) \\ &+ \frac{1}{n} q_2\left(\frac{x}{\sigma_n} \,\middle|\, P\right)\phi\left(\frac{x}{\sigma_n}\right) + \cdots.\end{aligned}$$

The functions q_i are polynomials, which are generally different from the polynomials occurring in the Edgeworth expansion for the studentized statistic. The bootstrap version of this expansion is

$$\begin{aligned}\mathrm{P}(\hat{\theta}_n^* - \hat{\theta}_n \leq x \mid \hat{P}_n) = \Phi\left(\frac{x}{\hat{\sigma}_n}\right) &+ \frac{1}{\sqrt{n}} q_1\left(\frac{x}{\hat{\sigma}_n} \,\middle|\, \hat{P}_n\right)\phi\left(\frac{x}{\hat{\sigma}_n}\right) \\ &+ \frac{1}{n} q_2\left(\frac{x}{\hat{\sigma}_n} \,\middle|\, \hat{P}_n\right)\phi\left(\frac{x}{\hat{\sigma}_n}\right) + \cdots.\end{aligned}$$

The Kolmogorov-Smirnov distance between the distribution functions on the left in the preceding displays is of the same order as the difference between the leading terms $\Phi(x/\sigma_n) - \Phi(x/\hat{\sigma}_n)$ on the right. Because the estimator $\hat{\sigma}_n$ is typically not closer than $O_P(n^{-1/2})$ to σ, this difference may be expected to be at best of the order $O_P(n^{-1/2})$. Thus, the percentile method for estimating a distribution is correct only up to the order $O_P(n^{-1/2})$, whereas the percentile t-method is seen to be correct up to the order $O_P(n^{-1})$.

One-sided bootstrap percentile t and percentile confidence intervals attain orders of correctness that are equal to the orders of correctness of the bootstrap estimators of the distribution functions: $O_P(n^{-1})$ and $O_P(n^{-1/2})$, respectively. For equal-tailed confidence intervals both methods typically have coverage error of the order $O_P(n^{-1})$. The decrease in coverage error is due to the cancellation of the errors contributed by the left and right tails, just as in the case of normal confidence intervals. The proofs of these assertions are somewhat technical. The coverage probabilities can be expressed in probabilities of the type

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le \hat{\xi}_{n,\alpha} \mid P\right). \tag{23.15}$$

Thus we need an Edgeworth expansion of the distribution of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n - \hat{\xi}_{n,\alpha}$, or a related quantity. A technical complication is that the random variables $\hat{\xi}_{n,\alpha}$ are only implicitly defined, as the solution of (23.1).

To find the expansions, first evaluate the Edgeworth expansion for $(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ at its the upper quantile $\hat{\xi}_{n,\alpha}$ to find that

$$1 - \alpha = \Phi(\hat{\xi}_{n,\alpha}) + \frac{p_1(\hat{\xi}_{n,\alpha} \mid \hat{P}_n)\phi(\hat{\xi}_{n,\alpha})}{\sqrt{n}} + O_P\left(\frac{1}{n}\right).$$

After expanding Φ, p_1 and ϕ in Taylor series around z_α, we can invert this equation to obtain the (conditional) *Cornish-Fisher expansion*

$$\hat{\xi}_{n,\alpha} = z_\alpha - \frac{p_1(z_\alpha \mid P)}{\sqrt{n}} + O_P\left(\frac{1}{n}\right).$$

In general, Cornish-Fisher expansions are asymptotic expansions of quantile functions, much in the same spirit as Edgeworth expansions are expansions of distribution functions. The probability (23.15) can be rewritten

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} - O_P\left(\frac{1}{n}\right) \le z_\alpha - \frac{p_1(z_\alpha \mid P)}{\sqrt{n}} \,\middle|\, P\right).$$

For a rigorous derivation it is necessary to characterize the $O_P(n^{-1})$ term. Informally, this term should only contribute to terms of order $O(n^{-1})$ in an Edgeworth expansion. If we just ignore it, then the probability in the preceding display can be expanded with the help of (23.12) as

$$\Phi\left(z_\alpha - \frac{p_1(z_\alpha \mid P)}{\sqrt{n}}\right) + \frac{p_1\left(z_\alpha - n^{-1/2}p_1(z_\alpha \mid P) \mid P\right)}{\sqrt{n}}\phi\left(z_\alpha - \frac{p_1(z_\alpha \mid P)}{\sqrt{n}}\right) + O\left(\frac{1}{n}\right).$$

The linear term of the Taylor expansion of Φ cancels the leading term of the Taylor expansion of the middle term. Thus the expression in the last display is equal to $1 - \alpha$ up to the order

$O(n^{-1})$, whence the coverage error of a percentile t–confidence interval is of the order $O(n^{-1})$.

For percentile intervals we proceed in the same manner, this time inverting the Edgeworth expansion of the unstudentized statistic. The (conditional) Cornish-Fisher expansion for the quantile $\hat{\xi}_{n,\alpha}$ of $\hat{\theta}_n^* - \hat{\theta}_n$ takes the form

$$\frac{\hat{\xi}_{n,\alpha}}{\hat{\sigma}_n} = z_\alpha - \frac{q_1(z_\alpha \mid \hat{P}_n)}{\sqrt{n}} + O_P\left(\frac{1}{n}\right).$$

The coverage probabilities of percentile confidence intervals can be expressed in probabilities of the type

$$\mathrm{P}(\hat{\theta}_n - \theta \le \hat{\xi}_{n,\alpha} \mid P) = \mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le \frac{\hat{\xi}_{n,\alpha}}{\hat{\sigma}_n} \mid P\right).$$

Insert the Cornish-Fisher expansion, again neglect the $O_P(n^{-1})$ term, and use the Edgeworth expansion (23.12) to rewrite this as

$$\Phi\left(z_\alpha - \frac{q_1(z_\alpha \mid P)}{\sqrt{n}}\right) + \frac{p_1\left(z_\alpha - n^{-1/2}q_1(z_\alpha \mid P) \mid P\right)}{\sqrt{n}}\phi\left(z_\alpha - \frac{q_1(z_\alpha \mid P)}{\sqrt{n}}\right) + O\left(\frac{1}{n}\right).$$

Because p_1 and q_1 are different, the cancellation that was found for the percentile t-method does not occur, and this is generally equal to $1-\alpha$ up to the order $O(n^{-1/2})$. Consequently, asymmetric percentile intervals have coverage error of the order $O(n^{-1/2})$. On the other hand, the coverage probability of the symmetric confidence interval $[\hat{\theta}_n - \hat{\xi}_{n,\alpha}, \hat{\theta}_n - \hat{\xi}_{n,1-\alpha}]$ is equal to the expression in the preceding display minus this expression evaluated for $1-\alpha$ instead of α. In the common situation that both polynomials p_1 and q_1 are even, the terms of order $O(n^{-1/2})$ cancel, and the difference is equal to $1-2\alpha$ up to the order $O(n^{-1})$. Then the percentile two-tailed confidence interval has the same order of correctness as the symmetric normal interval and the percentile t-intervals.

Notes

For a wider scope on the applications of the bootstrap, see the book [44], whose first author Efron is the inventor of the bootstrap. Hall [72] gives a detailed treatment of higher-order expansions of a number of bootstrap schemes. For more information concerning the consistency of the empirical bootstrap, and the consistency of the bootstrap under the application of the delta method, see Chapter 3.6 and Section 3.9.3 of [146], or the paper by Giné and Zinn [58].

PROBLEMS

1. Let $\hat{F}_n$ be a sequence of random distribution functions and F a continuous, fixed-distribution function. Show that the following statements are equivalent:

(i) $\hat{F}_n(x) \xrightarrow{\mathrm{P}} F(x)$ for every x.

(ii) $\sup_x \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{\mathrm{P}} 0$.

2. Compare in a simulation study Efron's percentile method, the normal approximation in combination with Fisher's transformation, and the percentile method to set a confidence interval for the correlation coefficient.
3. Let $X_{(n)}$ be the maximum of a sample of size n from the uniform distribution on $[0, 1]$, and let $X^*_{(n)}$ be the maximum of a sample of size n from the empirical distribution $\mathbb{P}_n$ of the first sample. Show that $\mathrm{P}(X^*_{(n)} = X_{(n)} \mid \mathbb{P}_n) \to 1 - e^{-1}$. What does this mean regarding the consistency of the empirical bootstrap estimator of the distribution of the maximum?
4. Devise a bootstrap scheme for setting confidence intervals for β in the linear regression model $Y_i = \alpha + \beta x_i + e_i$. Show consistency.
5. **(Parametric bootstrap.)** Let $\hat{\theta}_n$ be an estimator based on observations from a parametric model P_θ such that $\sqrt{n}(\hat{\theta}_n - \theta - h_n/\sqrt{n})$ converges under $\theta + h_n/\sqrt{n}$ to a continuous distribution L_θ for every converging sequence h_n and every θ. (This is slightly stronger than *regularity* as defined in the chapter on asymptotic efficiency.) Show that the *parametric bootstrap* is consistent: If $\hat{\theta}^*_n$ is $\hat{\theta}_n$ computed from observations obtained from $P_{\hat{\theta}_n}$, then $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n) \rightsquigarrow L_\theta$ conditionally on the original observations, in probability. (The conditional law of $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)$ is $L_{n,\hat{\theta}}$ if $L_{n,\theta}$ is the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ under θ.)
6. Suppose that $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow T$ and $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n) \rightsquigarrow T$ in probability given the original observations. Show that $\sqrt{n}\big(\phi(\hat{\theta}^*_n) - \phi(\hat{\theta}_n)\big) \rightsquigarrow \phi'_\theta(T)$ in probability for every map ϕ that is differentiable at θ.
7. Let U_n be a U-statistic based on a random sample $X_1, \ldots, X_n$ with kernel $h(x, y)$ such that both $\mathrm{E}h(X_1, X_1)$ and $\mathrm{E}h^2(X_1, X_2)$ are finite. Let $\hat{U}^*_n$ be the same U-statistic based on a sample $X^*_1, \ldots, X^*_n$ from the empirical distribution of $X_1, \ldots, X_n$. Show that $\sqrt{n}(\hat{U}^*_n - U_n)$ converges conditionally in distribution to the same limit as $\sqrt{n}(U_n - \theta)$, almost surely.
8. Suppose that $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow T$ and $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n) \rightsquigarrow T$ in probability given the original observations. Show that, unconditionally, $\sqrt{n}(\hat{\theta}_n - \theta, \hat{\theta}^*_n - \hat{\theta}_n) \rightsquigarrow (S, T)$ for independent copies S and T of T. Deduce the unconditional limit distribution of $\sqrt{n}(\hat{\theta}^*_n - \theta)$.