METHODOLOGY AND THEORY FOR THE BOOTSTRAP

Peter Hall

1. **Bootstrap principle:** definition; history; examples of problems that can be solved; different versions of the bootstrap

2. **Explaining the bootstrap in theoretical terms:** introduction to (Chebyshev-)Edgeworth approximations to distributions; rigorous development of Edgeworth expansions; 'smooth function model'; Edgeworth-based explanations for the bootstrap

3. **Bootstrap iteration:** principle and theory

4. **Bootstrap in non-regular cases:** Difficulties that the bootstrap has modelling extremes; m-out-of-n bootstrap; bootstrap for curve estimation

5. **Bootstrap for time series:** 'structural' and 'non-structural' implementations; block bootstrap methods

6. **Speeding the performance of Monte Carlo simulation**

What is the bootstrap?

The bootstrap is the statistical procedure which models sampling from a population by the process of resampling from the sample.

Therefore, if a quantity (e.g. a parameter) can be expressed as a functional of an unknown distribution, then its bootstrap estimator is the same functional of the empirical distribution function.

Simple examples

Best known example is the case of the mean, μ say, of a population drawn by sampling randomly from a population with distribution function F:

$$\mu = \int x \, dF(x) \, .$$

Its mean is the same functional of the empirical distribution function F_n, i.e. of

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x) \, ,$$

where $X_1, \ldots, X_n$ denote the data. Therefore the bootstrap estimator of the population mean, μ, is the sample mean, $\bar{X}$:

$$\bar{X} = \int x \, d\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

Likewise, the bootstrap estimator of a population variance is the corresponding sample variance; the bootstrap estimator of a population correlation coefficient is the corresponding empirical correlation coefficient; and so on.

Note particularly that Monte Carlo simulation does not play a role in the definition of the bootstrap, although simulation is an essential feature of most implementations of bootstrap methods.

Prehistory of the bootstrap

In view of the definition above, one could fairly argue that the calculation and application of bootstrap estimators has been with us for centuries.

One could claim that general first-order limit theory for the bootstrap was known to Laplace by about 1810 (since Laplace developed one of the earliest general central limit theorems); and that second-order properties were developed by Chebyshev at the end of the 19th Century. (Chebyshev was one of the first to explore properties of what we usually refer to today as *Edgeworth expansions*.)

However, a ‘mathematical’ or ‘technical’ approach to defining the bootstrap, and hence to defining its history, tends to overlook its most important feature: using sampling from the sample to model sampling from the population.

Sample surveys and the bootstrap

The notion of sampling from a sample is removed only slightly from that of sampling from a finite population. Unsurprisingly, then, a strong argument can be made that important aspects of the bootstrap's roots lie in methods for sample surveys.

There, the variance of samples drawn from a sample have long been use used to assess sampling variability, and to assess sampling variation.

J.A. Hubback

Arguably the first person to be involved in this type of work was not a statistician but an Indian Civil Servant, John Hubback. Hubback, and Englishman, was born in 1878 and worked in India for most of the 45 year period after 1902. He died in 1968.

In 1923 Hubback began a series of crop trials, in the Indian states of Bihar and Orissa, in which he developed spatial sampling schemes. In 1927 he published an account of his work in a Bulletin of the Indian Agricultural Research Institute.

Hubback went on to become the first governor of Orissa province. As Sir John Hubback he served as an advisor to Lord Mountbatten's administration of India, at the end of British rule.

J.A. Hubback (continued)

Hubback's work was to have a substantial influence on subsequent work on random sampling for assessing crop yields in the UK, conducted at Rothamsted by Fisher and Yates. Fisher was to write:

The use of the method of random sampling is theoretically sound. I may mention that its practicability, convenience and economy was demonstrated by an extensive series of crop-cutting experiments on paddy carried out by Hubback.... They influenced greatly the development of my methods at Rothamsted. (R.A. Fisher, 1945)

P.C. Mahalanobis

Mahalanobis, the eminent Indian statistician, was inspired by Hubback's work and used Hubback's spatial sampling schemes explicitly for variance estimation. This was a true precursor of bootstrap methods.

Of course, Mahalanobis appreciated that the data he was sampling were correlated, and he carefully assessed the effects of dependence, both empirically and theoretically. His work in the late 1930s, and during the War, anticipated the much more modern technique of the block bootstrap.

Other contributors

So-called ‘half-sampling’ methods were used by the US Bureau of the Census from at least the late 1950s. This pseudo-replication technique was designed to produce, for stratified data, an effective estimator of the variance of the grand mean (a weighted average over strata) of the data. The aim was to improve on the conventional variance estimator, computed as a weighted linear combination of within-stratum sample variances.

Names associated with methodological development of half-sampling include Gurney (1962) and McCarthy (1966, 1969). Substantial contributions on the theoretical side were made by Hartigan (1969, 1971, 1975).

Julian Simon, and others

Permutation methods related to the bootstrap were discussed by Maritz (1978) and Maritz and Jarrett (1978), and by the Social Scientist Julian Simon, who wrote as early as 1969 that computer-based experimentation in statistics “holds great promise for the future.”

Unhappily, Simon (who died in 1998) spent a significant part of the 1990s disputing with some of the statistics profession his claims to have ‘discovered’ the bootstrap. He argued that statisticians had only grudgingly accepted ‘his’ ideas on the bootstrap, and and borrowed them without appropriate attribution.

Julian Simon (continued)

Simon saw the community of statisticians as an unhappy 'priesthood', which felt jealous because the computer-based bootstrap made their mathematical skills redundant:

The simple fact is that resampling devalues the knowledge of conventional mathematical statisticians, and especially the less competent ones. By making it possible for each user to develop her/his own method to handle each particular problem, the priesthood with its secret formulaic methods is rendered unnecessary. No one...stands still for being rendered unnecessary. Instead, they employ every possible device fair and foul to repel the threat to the economic well-being and their self-esteem.

Efron’s contributions

Efron’s contributions, the ramifications of which we shall explore in subsequent lectures, were of course far-reaching. They vaulted forward from the earlier ideas, of people such as Hubback, Mahalanobis, Hartigan and Simon, creating a fully fledged methodology that is now applied to analyse data on virtually all human beings (e.g. through the bootstrap for sample surveys).

Efron married the power of Monte Carlo approximation with an exceptionally broad view of the sort problem that bootstrap methods might solve. For example, he saw that the notion of a ‘parameter’ (that functional of a distribution function, which we considered earlier) might be interpreted very widely, and taken to be (say) the coverage level of a confidence interval.

Main principle

Many statistical problems can be represented as follows: given a functional f_t from a class $\{f_t\colon\ t \in \mathcal{T}\}$, we wish to determine the value of a parameter t that solves an equation,

$$E\{f_t(F_0, F_1) \mid F_0\} = 0\,, \tag{1}$$

where F_0 denotes the population distribution function and F_1 is the distribution function 'of the sample' — that is, the empirical distribution function $\widehat{F}$.

Example 1: bias correction

Here, $\theta = \theta(F_0)$ is the true value of a parameter, and $\widehat{\theta} = \theta(F_1)$ is its estimator; t is an additive adjustment to $\widehat{\theta}$; $\widehat{\theta} + t$ is the bias-corrected estimator; and

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t$$

denotes the bias-corrected version of $\widehat{\theta}$, minus the true value of the parameter. Ideally, we would like to choose t so as to reduce bias to zero, i.e. so as to solve $E(\widehat{\theta} - \theta + t) = 0$, which is equivalent to (1).

Example 2: confidence interval

Here we take

$$\begin{aligned} f_t(F_0, F_1) &= I\left\{\theta(F_1) - t \le \theta(F_0) \le \theta(F_1) + t\right\} \\ &\quad - (1 - \alpha), \end{aligned}$$

denoting the indicator of the event that the true parameter value $\theta(F_0)$ lies in the interval,

$$[\theta(F_1) - t, \theta(F_1) + t] = [\widehat{\theta} - t, \widehat{\theta} + t],$$

minus the nominal coverage, $1 - \alpha$, of the interval. (Thus, the chosen interval is two-sided and symmetric.) Asking that

$$E\{f_t(F_0, F_1) \mid F_0\} = 0$$

is equivalent to insisting that t be chosen so that the interval has zero coverage error.

Bootstrapping equation (1)

We call equation (1), i.e.

$$E\{f_t(F_0, F_1) \mid F_0\} = 0\,, \tag{1}$$

the *population equation*. The *sample equation* is obtained by replacing the pair (F_0, F_1) by (F_1, F_2), where $F_2 = \widehat{F}^*$ is the bootstrap form of the empirical distribution function F_1:

$$E\{f_t(F_1, F_2) \mid F_1\} = 0\,. \tag{2}$$

Recall that

$$F_1(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)\,;$$

analogously, we define

$$F_2(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i^* \leq x)\,,$$

where the bootstrap sample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ is obtained by sampling randomly, with replacement, from the original sample $\mathcal{X} = \{X_1, \ldots, X_n\}$.

Sampling randomly, with replacement

'Sampling randomly, with replacement from $\mathcal{X}$' means that

$$P\left(X_i^* = X_j \mid \mathcal{X}\right) = \frac{1}{n}$$

for $i, j = 1, \ldots, n$.

This is standard 'random, uniform bootstrap sampling.' More generally, we might *tilt* the empirical distribution $F_1 = \widehat{F}$ by sampling with weight p_j attached to data value X_j:

$$P\left(X_i^* = X_j \mid \mathcal{X}\right) = p_j$$

for $i, j = 1, \ldots, n$. Of course, we should insist that the p_i's form a multinomial distribution, i.e. satisfy $p_i \geq 0$ and $\sum_i p_i = 1$.

Tilting is used in many contemporary generalisations of the bootstrap, such as empirical likelihood and the weighted, or biased bootstrap.

Example 1, revisited: bias correction

Recall that the population and sample equations are here given by

$$\begin{aligned} E\{\theta(F_1)-\theta(F_0)+t \mid F_0\} &= 0\,, \\ E\{\theta(F_2)-\theta(F_1)+t \mid F_1\} &= 0\,, \end{aligned}$$

respectively. Clearly the solution of the latter is

$$\begin{aligned} t=\hat{t} &= \theta(F_1)-E\{\theta(F_2) \mid F_1\} \\ &= \hat{\theta}-E(\hat{\theta}^* \mid \hat{F})\,. \end{aligned}$$

This is the bootstrap estimate of the additive correction that should be made to $\hat{\theta}$ in order to reduce bias. The bootstrap bias-corrected estimator is thus

$$\hat{\theta}_{\text{bc}}=\hat{\theta}+\hat{t}=2\,\hat{\theta}-E(\hat{\theta}^* \mid \hat{F})\,,$$

where the subscript bc denotes ‘bias corrected.’

Example 2, revisited: confidence interval

In the confidence-interval example, the sample equation has the form

$$P\{\theta(F_2) - t \le \theta(F_1) \le \theta(F_2) + t \mid F_1\} \\ -(1-\alpha) = 0\,,$$

or equivalently,

$$P(\widehat{\theta}^* - t \le \widehat{\theta} \le \widehat{\theta}^* + t \mid \mathcal{X}) = 1 - \alpha\,.$$

Since $\widehat{\theta}$, conditional on $\mathcal{X}$, has a discrete distribution then it is seldom possible to solve exactly for t. However, any error is usually small, since the size of even the largest atom decreases exponentially fast with increasing n.

We could remove this difficulty by smoothing the distribution F_1, and this is sometimes done in practice.

Example 2, revisited (continued; 1)

To obtain an approximate solution, $t = \hat{t}$, of the equation

$$P(\widehat{\theta}^* - t \leq \widehat{\theta} \leq \widehat{\theta}^* + t \mid \mathcal{X}) = 1 - \alpha\,.$$

we use Monte Carlo methods. That is, conditional on $\mathcal{X}$ we calculate independent values $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ of $\widehat{\theta}^*$, and take $\hat{t}(B)$ to be an approximate solution of the equation

$$\frac{1}{B} \sum_{b=1}^{B} I(\widehat{\theta}_b^* - t \leq \widehat{\theta} \leq \widehat{\theta}_b^* + t) = 1 - \alpha\,.$$

For example, it might denote the largest t such that

$$\frac{1}{B} \sum_{b=1}^{B} I(\widehat{\theta}_b^* - t \leq \widehat{\theta} \leq \widehat{\theta}_b^* + t) \leq 1 - \alpha\,.$$

Example 2, revisited (continued; 2)

The resulting confidence interval is a standard 'percentile method' bootstrap confidence interval for θ. Under mild regularity conditions its limiting coverage, as $n \to \infty$, is $1 - \alpha$, and its coverage error equals $O(n^{-1})$. That is,

$$P(\widehat{\theta} - \widehat{t} \le \theta \le \widehat{\theta} + \widehat{t}) = 1 - \alpha + O(n^{-1})\,. \quad (3)$$

Interestingly, this result is hardly affected by the number of bootstrap simulations we do. Usually one derives (3) under the assumption that $B = \infty$, but it can be shown that (3) remains true uniformly in $B_0 \le B \le \infty$, for finite B. However, we need to make a minor change to the way we construct the interval, which we shall discuss shortly in the case of two-sided intervals.

Example 2, revisited (continued; 3)

As we shall see later, the good coverage accuracy of two-sided intervals is the result of fortuitous cancellation of terms in approximations to coverage error (Edgeworth expansions). No such cancellation occurs in the case of one-sided versions of percentile confidence intervals, for which coverage error is generally only $O(n^{-1/2})$ as $n \to \infty$.

A one-sided percentile confidence interval for θ is given by $(-\infty, \widehat{\theta} + \widehat{t}]$, where $t = \widehat{t}$ is the (approximate) solution of the equation

$$P(\widehat{\theta} \leq \widehat{\theta}^* + t \mid \mathcal{X}) = 1 - \alpha\,.$$

Example 2, revisited (continued; 4)

(Here we explain how to construct a one-sided interval so that its coverage performance is not adversely affected by too-small choice of B.) Observing that B simulated values of $\widehat{\theta}$ divide the real line into $B+1$ parts, choose B, and an integer ν, such that

$$\frac{\nu}{B+1} = 1 - \alpha\,. \tag{4}$$

(For example, in the case $\alpha = 0.05$ we might take $B = \nu = 19$.) Let $\widehat{\theta}_{(\nu)}$ denote the νth largest of the B simulated values of $\widehat{\theta}^*$, and let the confidence interval be $(-\infty, \widehat{\theta}_{(\nu)}]$. Then,

$$P\left\{\theta \in (-\infty, \widehat{\theta}_{(\nu)}]\right\} = \alpha + O(n^{-1/2})$$

uniformly in pairs (B, ν) such that (4) holds, as $n \to \infty$.

Combinatorial calculations connected with the bootstrap

- If the sample $\mathcal{X}$ is of size n, and if all its elements are distinct, then the number, $N(n)$ say, of different possible resamples $\mathcal{X}$ that can be drawn equals the number of ways of placing n indistinguishable objects into n numbered boxes (box i representing X_i), the boxes being allowed to contain any number of objects. (The number, m_i say, of objects in box i represents the number of times X_i appears in the sample.)

- In fact, $N(n) = \binom{2n-1}{n}$. (Exercise: prove this!) Therefore, the bootstrap distribution, for a sample of n distinguishable data, has just $\binom{2n-1}{n}$ atoms.

Combinatorial calculations (continued; 1)

- The value of $N(n)$ increases exponentially fast with n; indeed, $N(n) \sim (n\pi)^{-1/2} 2^{2n-1}$.

n	$N(n)$
2	3
3	10
4	35
5	126
6	462
7	1716
8	6435
9	24310
10	92378
15	7.8×10^7
20	6.9×10^{10}

Combinatorial calculations (continued; 2)

- Not all the $N(n)$ atoms of the bootstrap distribution have equal mass. The most likely atom is that which arises when $\mathcal{X}^* = \mathcal{X}$, i.e. when the resample is identical to the full sample. Its probability: $p_n = n!/n^n \sim (2n\pi)^{1/2}e^{-n}$.

n	p_n
2	0.5
3	0.2222
4	0.0940
5	0.0384
6	1.5×10^{-2}
7	6.1×10^{-3}
8	2.4×10^{-3}
9	9.4×10^{-4}
10	3.6×10^{-4}
15	3.0×10^{-6}
20	2.3×10^{-8}

METHODOLOGY AND THEORY FOR THE BOOTSTRAP
(Second two lectures)

Main principle (revision)

We argued that many statistical problems can be represented as follows: given a functional f_t from a class $\{f_t\colon\ t \in \mathcal{T}\}$, we wish to determine the value of a parameter t that solves the *population equation*,

$$E\{f_t(F_0, F_1) \mid F_0\} = 0\,, \tag{1}$$

where F_0 denotes the population distribution function, and

$$F_1(x) = \widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x)$$

is the empirical distribution function, computed from the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$.

Let $t_0 = T(F_0)$ denote the solution of (1).

Revision (continued, 1)

We introduced a bootstrap approach to estimating t_0: solve instead the *sample equation*,

$$E\{f_t(F_1, F_2) \mid F_1\} = 0 \,, \qquad (2)$$

where

$$F_2(x) = \widehat{F}^*(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i^* \leq x)$$

is the bootstrap form of the empirical distribution function. (The bootstrap sample is $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$, drawn by sampling randomly, with replacement, from $\mathcal{X}$).)

The solution, $\hat{t} = T(F_1)$ say, of (2) is an estimator of the solution $t_0 = T(F_0)$ of (1). It does not itself solve (1), but (1) is usually approximately correct if $T(F_0)$ is replaced by $T(F_1)$:

$$E\{f_{T(F_1)}(F_0, F_1) \mid F_0\} \approx 0 \,.$$

How accurate is this approximation, and what does it mean?

We considered two examples, one of bias correction and the other of confidence intervals. In the bias-correction example,

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t = \widehat{\theta} - \theta_0 + t\,,$$

and here it is generally true that the error in the approximation is of order n^{-2}:

$$E\{f_{T(F_1)}(F_0, F_1) \mid F_0\} = O(n^{-2})\,.$$

Equivalently, the amount of uncorrected bias is of order n^{-2}: writing $\widehat{t}$ for $T(F_1)$,

$$E(\widehat{\theta} - \theta_0 + \widehat{t}) = O(n^{-2})\,.$$

That is an improvement on the amount of bias without any attempt at correction; this is usually only $O(n^{-1})$:

$$E(\widehat{\theta} - \theta_0) = O(n^{-1})\,.$$

How accurate is the approximation? (continued, 1)

The second example was of two-sided confidence intervals, and there,

$$\begin{aligned} f_t(F_0, F_1) \;&=\; I\left\{\theta(F_1) - t \le \theta(F_0) \le \theta(F_1) + t\right\} \\ &\qquad\qquad - \;(1-\alpha)\,, \end{aligned}$$

denoting the indicator of the event that the true parameter value $\theta(F_0)$ lies in the interval

$$[\theta(F_1) - t, \theta(F_1) + t] = [\widehat{\theta} - t, \widehat{\theta} + t]\,,$$

minus the nominal coverage, $1 - \alpha$, of the interval.

Solving the sample equation, we obtain an estimator, $\widehat{t}$, of the solution of the population equation. The resulting confidence interval,

$$[\widehat{\theta} - \widehat{t}, \widehat{\theta} + \widehat{t}]\,,$$

is generally called a *percentile* bootstrap confidence interval for θ, with nominal coverage $1 - \alpha$.

How accurate is the approximation? (continued, 2)

In this setting the error in the approximation to the population equation, offered by the sample equation, is usually of order n^{-1}. This time it means that the amount of uncorrected coverage error is of order n^{-1}:

$$\begin{aligned} P\{\theta(F_1) - \hat{t} \le \theta(F_0) \le \theta(F_1) + \hat{t}\} \\ = \ 1 - \alpha + O(n^{-1}) \, . \end{aligned}$$

That is,

$$P\{\hat{\theta} - \hat{t} \le \theta_0 \le \hat{\theta} + \hat{t}\} = 1 - \alpha + O(n^{-1}) \, .$$

Put another way, "the coverage error of the nominal $1-\alpha$ level, two-sided percentile bootstrap confidence interval $[\hat{\theta} - \hat{t}, \hat{\theta} + \hat{t}]$, equals $O(n^{-1})$."

How accurate is the approximation? (continued, 3)

However, coverage error in the one-sided case is usually only $O(n^{-1/2})$. That is, if we define $t = T(F_1) = \hat{t}$ to solve the population equation with

$$f_t(F_0, F_1) = I\left\{\theta(F_0) \le \theta(F_1) + t\right\} - (1 - \alpha)\,,$$

then

$$P\{\theta_0 \le \hat{\theta} + \hat{t}\} = 1 - \alpha + O(n^{-1/2})\,.$$

That is, "the coverage error of the nominal $1-\alpha$ level, one-sided percentile bootstrap confidence interval $(-\infty, \hat{\theta} + \hat{t}]$ equals $O(n^{-1/2})$."

Bootstrap iteration

Here we suggest iterating the "bootstrap principle" so as to produce a more accurate solution of the population equation.

Our solution currently has the property

$$E\{f_{T(F_1)}(F_0, F_1) \mid F_0\} \approx 0\,. \tag{3}$$

Let us replace $T(F_1)$ by a perturbation, which might be additive, $U(F_1, t) = T(F_1) + t$, or multiplicative, $U(F_1, t) = (1 + t)\, T(F_1)$. Substitute this for $T(F_1)$ in (1), and attempt to solve the resulting equation for t:

$$E\{f_{U(F_1,t)}(F_0, F_1) \mid F_0\} = 0\,.$$

This is no more than a re-writing of the original population equation, with a new definition of f. Our way of solving it will be the same as before — write down its sample version,

$$E\{f_{U(F_2,t)}(F_1, F_2) \mid F_1\} = 0\,, \tag{4}$$

and solve that.

Repeating bootstrap iteration

Of course, we can repeat this procedure as often as we wish.

Recall, however, that in most instances the sample equation can only be solved by Monte Carlo simulation: calculating $\widehat{t}$ involves drawing B resamples $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the original sample, $\mathcal{X} = \{X_1, \ldots, X_n\}$, by sampling randomly, with replacement. When solving the new sample equation,

$$E\{f_{U(F_2,t)}(F_1, F_2) \mid F_1\} = 0\,, \tag{4}$$

we have to sample from the resample. That is, in order to compute the solution of (4), from each given $\mathcal{X}^*$ in the original bootstrap resampling step we must draw data $X_1^{**}, \ldots,$ X_n^{**} by sampling randomly, with replacement; and combine these into a bootstrap re-resample $\mathcal{X}^{**} = \{X_1^{**}, \ldots, X_n^{**}\}$.

The computational expense of this procedure usually prevents more than one iteration.

Implementing the double bootstrap

Let us work through the example of one-sided bootstrap confidence intervals. Here, we ideally want t such that

$$P(\theta \le \widehat{\theta} + t) = 1 - \alpha\,,$$

where $1-\alpha$ is the nominal coverage level of the confidence interval. Our one-sided confidence interval for θ would then be $(-\infty, \widehat{\theta} + t)$.

One application of the bootstrap involves creating resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_n^*$; computing the version, $\widehat{\theta}_b^*$, of $\widehat{\theta}$ from $\mathcal{X}_b^*$; and choosing $t = \widehat{t}$ such that

$$\frac{1}{B} \sum_{b=1}^{B} I(\widehat{\theta} \le \widehat{\theta}_b^* + t) = 1 - \alpha\,,$$

where we solve the equation as nearly as possible. (We do not actually use this $\widehat{t}$ for the iterated, or double, bootstrap step, but it gives us the standard bootstrap percentile confidence interval $(-\infty, \widehat{\theta} + \widehat{t})$.)

Implementing the double bootstrap (continued, 1)

For the next application of the bootstrap, from each resample $\mathcal{X}_b^*$ draw C re-resamples, $\mathcal{X}_{b1}^{**}$, $\ldots, \mathcal{X}_{bC}^{**}$, the cth (for $1 \le c \le C$) given by

$$\mathcal{X}_{bc}^{**} = \{X_{bc1}^{**}, \ldots, X_{bcn}^{**}\} ;$$

$\mathcal{X}_{bc}^{**}$ is obtained by sampling randomly, with replacement, from $\mathcal{X}_b^*$. Compute the version, $\widehat{\theta}_{bc}^{**}$, of $\widehat{\theta}$ from $\mathcal{X}_{bc}^{**}$; and choose $t = \widehat{t}_b^*$ such that

$$\frac{1}{C} \sum_{c=1}^{C} I(\widehat{\theta}_b^* \le \widehat{\theta}_{bc}^{**} + t) = 1 - \alpha \, ,$$

as nearly as possible.

Implementing the double bootstrap (continued, 2)

Interpret $\widehat{t}_b^*$ as the version of $\widehat{t}$ we would employ if the sample were $\mathcal{X}_b^*$, rather than $\mathcal{X}$. We "calibrate" or "correct" it, using the perturbation argument introduced earlier.

Let us take the perturbation to be additive, for definiteness. Then we find $t = \tilde{t}$ such that

$$\frac{1}{B} \sum_{b=1}^{B} I(\widehat{\theta} \leq \widehat{\theta}_b^* + \widehat{t}_b^* + t) = 1 - \alpha \,,$$

as nearly as possible.

Our final double-bootstrap, or bootstrap-calibrated, one-sided percentile confidence interval is

$$(-\infty, \widehat{\theta} + \widehat{t} + \tilde{t}] \,.$$

How successful is bootstrap iteration?

Each application of bootstrap iteration usually improves the order of accuracy by an order of magnitude.

For example, in the case of bias correction each application generally reduces the order of bias by a factor of n^{-1}.

In the case of one-sided confidence intervals, each application usually reduces the order of coverage error by the factor $n^{-1/2}$. Recall that the standard percentile bootstrap confidence interval has coverage error $n^{-1/2}$. Therefore, applying one iteration of the bootstrap (i.e. the double bootstrap) reduces the order of error to $n^{-1/2} \times n^{-1/2} = n^{-1}$.

Shortly we shall see that it is possible to construct uncalibrated bootstrap one-sided confidence intervals that have coverage error n^{-1}. Application of the double bootstrap to them reduces the order of their coverage error to $n^{-1/2} \times n^{-1} = n^{-3/2}$.

How successful is bootstrap iteration? (continued)

In the case of two-sided confidence intervals, each application usually reduces the order of coverage error by the factor n^{-1}. The standard percentile bootstrap confidence interval has coverage error n^{-1}, and after applying the double bootstrap this reduces to n^{-2}.

A subsequent iteration, if computationally feasible, would reduce coverage error to n^{-3}.

Note on choice of B and C

Recall that implementation of the double bootstrap is via two stages of bootstrap simulation, involving B and C simulations respectively. The total cost of implementation is proportional to BC. How should computational labour be distributed between the two stage?

A partial answer is that C should be of the same order as $\sqrt{B}$. As this implies, a high degree of accuracy in the second stage is less important than for the first stage.

Iterated bootstrap for bias correction

By its nature, the case of bias correction is relatively amenable to analytic treatment in general cases. We have already noted (in an earlier lecture) that the additive bootstrap bias adjustment, $\widehat{t} = T(F_1)$, is given by

$$T(F_1) = \theta(F_1) - E\{\theta(F_2) \mid F_1\}\,.$$

Therefore, the bias-corrected form of the estimator $\theta(F_1)$ is

$$\widehat{\theta}_1 = \theta(F_1) + T(F_1) = 2\,\theta(F_1) - E\{\theta(F_2) \mid F_1\}\,.$$

Iterated bootstrap for bias correction (continued, 1)

More generally, it may be proved by induction that, after j iterations of the bootstrap bias correction argument, we obtain the estimator $\widehat{\theta}_j$ given by

$$\widehat{\theta}_j = \sum_{i=1}^{j+1} \binom{j+1}{i} (-1)^{i+1} E\{\theta(F_i) \mid F_1\}\,. \quad (1)$$

Here F_i, for $i \geq 1$, denotes the empirical distribution function of a sample obtained by sampling randomly from the distribution F_{i-1}.

Exercise: Derive (1)*.*

Formula (1) makes explicitly clear the fact that, generally speaking, carrying out j bootstrap iterations involves computation of F_1, $\ldots, F_{j+1}$.

Iterated bootstrap for bias correction (continued, 2)

The bias of $\widehat{\theta}_j$ is generally of order $n^{-(j+1)}$; the original, non-iterated bootstrap estimator $\widehat{\theta}_0 = \widehat{\theta} = \theta(F_1)$ generally has bias of order n^{-1}.

Of course, there is a penalty to be paid for bias reduction: variance usually increases. However, asymptotic variance typically does not, since successive bias corrections are relatively small in size. Nevertheless, small-sample effects, on variance, of bias correction by bootstrap or other means are generally observable.

Iterated bootstrap for bias correction (continued, 3)

It is of interest to know the limit, as $j \to \infty$, of the estimator defined at (1). Provided $\theta(F)$ is an analytic function the limit can generally be worked out, and shown to be an unbiased estimator of θ with the same asymptotic variance as the original estimator $\widehat{\theta}$ (although larger variance in small samples).

Sometimes, but not always, the $j \to \infty$ limit is identical to the estimator obtained by a single application of the jackknife. Two elementary examples show this side of bootstrap bias correction.

Iterated bootstrap for bias correction (continued, 4)

Variance estimation

The conventional biased estimator of population variance, σ^2, is

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

whereas its unbiased form uses divisor $n-1$:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \,,$$

Noting that

$$\sigma^2(F_0) = \int x^2 \, dF_0(x) - \left\{ \int x \, dF_0(x) \right\}^2 \,,$$

we may write $\widehat{\sigma}^2$ in the usual bootstrap form, as $\widehat{\sigma}^2 = \sigma^2(\widehat{F})$. Therefore, $\widehat{\sigma}^2$ is the standard bootstrap variance estimator.

Iterating $\widehat{\sigma}_1 = \widehat{\sigma}^2$ through values $\widehat{\sigma}_j^2$, we find that as $j \to \infty$, $\widehat{\sigma}_j^2 \to S^2$. We can achieve the same limit in one step by using a multiplicative bias correction, or by the jackknife.

Percentile-t confidence intervals

The only bootstrap confidence intervals we have treated so far have been of the percentile type, where the interval endpoint is, in effect, a percentile of the bootstrap distribution.

In pre-bootstrap statistics, however, confidence regions were usually constructed very differently, using variance estimates and "Studentising," or pivoting, prior to using a central limit theorem to compute confidence limits.

These ideas have a role to play in the bootstrap case, too.

Percentile-t confidence intervals (continued, 1)

Let $\widehat{\theta}$ be an estimator of a parameter θ, and let $n^{-1}\widehat{\sigma}^2$ denote an estimator of its variance. In regular cases,

$$T = n^{1/2}(\widehat{\theta} - \theta)/\widehat{\sigma}$$

is asymptotically Normally distributed. In pre-bootstrap days one would have used this property to compute the approximate α-level quantile, t_α say, of the distribution of T, and used it to give a confidence interval for θ.

Specifically,

$$\begin{aligned} P(\theta \le \widehat{\theta} - n^{-1/2}t_\alpha) & \\ &= 1 - P\{n^{1/2}(\widehat{\theta} - \theta) \le t_\alpha\} \\ &\approx 1 - P\{\mathrm{N}(0,1) \le t_\alpha\} = 1 - \alpha\,, \end{aligned}$$

where the approximation derives from the central limit theorem. Hence, $(-\infty, \widehat{\theta} - n^{-1/2}t_\alpha]$ is an approximate $(1-\alpha)$-level confidence interval for θ.

Percentile-t confidence intervals (continued, 2)

Now we can improve on this approach by using the bootstrap, rather than the central limit theorem, to approximate the distribution of T.

Specifically, let $\widehat{\theta}^*$ and $\widehat{\sigma}^*$ denote the bootstrap versions of $\widehat{\theta}$ and $\widehat{\sigma}$ (i.e. the versions of $\widehat{\theta}$ and $\widehat{\sigma}$ computed from a resample $\mathcal{X}^*$, rather than the sample $\mathcal{X}$). Put

$$T^* = n^{1/2}(\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*\,,$$

and let $\widehat{t}_\alpha$ denote the α-level quantile of the bootstrap distribution of T^*:

$$P(T^* \leq \widehat{t}_\alpha \mid \mathcal{X}) = \alpha\,.$$

Percentile-t confidence intervals (continued, 3)

Recall that the Normal-approximation confidence interval for θ was

$$(-\infty, \widehat{\theta} - n^{-1/2} t_\alpha]\,.$$

If we replace t_α by its more accurate bootstrap form, $\widehat{\theta}_\alpha$, we obtain the percentile-t bootstrap confidence interval for θ:

$$(-\infty, \widehat{\theta} - n^{-1/2} \widehat{t}_\alpha]\,.$$

The coverage error of the former confidence is usually only $O(n^{-1/2})$:

$$P(\theta \leq \widehat{\theta} - n^{-1/2} t_\alpha) = \alpha + O(n^{-1/2})\,.$$

Likewise, the percentile-bootstrap confidence interval also has coverage error only of this size, unless we use the double bootstrap to calibrate it.

However, the percentile-t bootstrap confidence interval has coverage error $O(n^{-1})$:

$$P(\theta \leq \widehat{\theta} - n^{-1/2} \widehat{t}_\alpha) = \alpha + O(n^{-1})\,.$$

METHODOLOGY AND THEORY FOR THE BOOTSTRAP (Third set of two lectures)

Main topic of these lectures: Edgeworth expansions

Moments and cumulants

Let X be a random variable. Write $\chi(t) = E(e^{itX})$ for the associated characteristic function, and let κ_j denote the jth cumulant of X, i.e. the coefficient of $(it)^j/j!$ in an expansion of $\log \chi(t)$:

$$\begin{aligned}\chi(t) \; &= \; \exp\Big\{\kappa_1\, it + \tfrac{1}{2}\,\kappa_2\,(it)^2 + \ldots \\ &\qquad\qquad + \tfrac{1}{j!}\,\kappa_j\,(it)^j + \ldots\Big\}\,.\end{aligned}$$

The jth moment, $\mu_j = E(X^j)$, of X is the coefficient of $(it)^j/j!$ in an expansion of $\chi(t)$:

$$\begin{aligned}\chi(t) \; &= \; 1 + \mu_1\, it + \tfrac{1}{2}\,\mu_2\,(it)^2 + \ldots \\ &\qquad\qquad + \tfrac{1}{j!}\,\mu_j\,(it)^j + \ldots\end{aligned}$$

Expressing cumulants in terms of moments, and vice versa

Comparing these expansions we deduce that

$$\begin{aligned}
&\kappa_1 = \mu_1 \,, \quad \kappa_2 = \mu_2 - \mu_1^2 = \mathrm{var}(X)\,, \\
&\kappa_3 = \mu_3 - 3\,\mu_2\,\mu_1 + 2\,\mu_1^3 = E(X - EX)^3\,, \\
&\kappa_4 = \mu_4 - 4\,\mu_3\,\mu_1 - 3\,\mu_2^2 + 12\,\mu_2\,\mu_1^2 - 6\mu_1^4 \\
&\quad = E(X - EX)^4 - 3\,(\mathrm{var}X)^2\,.
\end{aligned}$$

In particular, κ_j is a homogeneous polynomial in moments, of degree j. Likewise, μ_j is a homogeneous polynomial in cumulants, of degree j.

Third and fourth cumulants, κ_3 and κ_4, are referred to as skewness and kurtosis, respectively.

Exercise: Express μ_j in terms of $\kappa_1, \ldots, \kappa_j$ for $j = 1, \ldots, 4$. Prove that, for $j \geq 2$, κ_j is invariant under translations of X.

Sums of independent random variables

Let us assume $\mu_1 = 0$ and $\mu_2 = 1$. This is equivalent to working with the normalised random variable $Y = (X - \mu_1)/\kappa_2^{1/2}$, instead of Y, although we shall continue to use the notation X rather than Y.

Let $X_1, X_2, \ldots$ be independent and identically distributed as X, and put

$$S_n = n^{-1/2} \sum_{j=1}^{n} X_j \,.$$

The characteristic function of S_n is

$$\begin{aligned}
\chi_n(t) &= E\,\{\exp(itS_n)\} \\
&= E\{\exp(itX_1/n^{1/2}) \ldots \exp(itX_n/n^{1/2})\} \\
&= E\{\exp(itX_1/n^{1/2})\} \ldots \\
&\qquad\qquad \times E\{\exp(itX_n/n^{1/2})\} \\
&= \chi(t/n^{1/2}) \ldots \chi(t/n^{1/2}) = \chi(t/n^{1/2})^n \,.
\end{aligned}$$

Sums of independent random variables (continued)

Therefore, since $\kappa_1 = 0$ and $\kappa_2 = 1$,

$$\begin{aligned}\chi_n(t) &= \chi(t/n^{1/2})^n \\ &= \Big[\exp\Big\{\kappa_1\,(it/n^{1/2}) + \ldots \\ &\qquad\qquad + \tfrac{1}{j!}\,\kappa_j\,(it/n^{1/2})^j + \ldots\Big\}\Big]^n \\ &= \exp\Big\{-\tfrac{1}{2}\,t^2 + n^{-1/2}\,\tfrac{1}{6}\,\kappa_3\,(it)^3 + \ldots \\ &\qquad\qquad + n^{-(j-2)/2}\,\tfrac{1}{j!}\,\kappa_j\,(it)^j + \ldots\Big\}\,.\end{aligned}$$

Now expand the exponent:

$$\begin{aligned}\chi_n(t) = e^{-t^2/2}\,\Big\{&1 + n^{-1/2}\,r_1(it) + \ldots \\ &+ n^{-j/2}\,r_j(it) + \ldots\Big\}\,,\end{aligned}$$

where r_j denotes a polynomial with real coefficients, of degree $3j$, having the same parity as its index, its coefficients depending on $\kappa_3, \ldots, \kappa_{j+2}$ but not on n. In particular,

$$r_1(u) = \tfrac{1}{6}\,\kappa_3\,u^3\,,\quad r_2(u) = \tfrac{1}{24}\,\kappa_4\,u^4 + \tfrac{1}{72}\,\kappa_3^2\,u^6\,.$$

Exercise: Prove this result, and the parity property of r_j.

Expansion of distribution function

Rewrite the expansion as:

$$\begin{aligned}\chi_n(t) = e^{-t^2/2} + n^{-1/2}\, r_1(it)\, e^{-t^2/2} + \ldots \\ +\ n^{-j/2}\, r_j(it)\, e^{-t^2/2} + \ldots .\end{aligned}$$

Note that

$$\begin{aligned}\chi_n(t) &= \int_{-\infty}^{\infty} e^{itx}\, dP(S_n \le x)\,, \\ e^{-t^2/2} &= \int_{-\infty}^{\infty} e^{itx}\, d\Phi(x)\,,\end{aligned}$$

where Φ denotes the standard Normal distribution function. Therefore, the expansion of $\chi_n(t)$ strongly suggests an "inverse" expansion,

$$\begin{aligned}P(S_n \le x) &= \Phi(x) + n^{-1/2}\, R_1(x) + \ldots \\ &\qquad + n^{-j/2}\, R_j(x) + \ldots\,,\end{aligned}$$

where

$$\int_{-\infty}^{\infty} e^{itx}\, dR_j(x) = r_j(it)\, e^{-t^2/2}\,.$$

Finding a formula for R_j

Integration by parts gives:

$$\begin{aligned} e^{-t^2/2} &= \int_{-\infty}^{\infty} e^{itx}\, d\Phi(x) \\ &= (-it)^{-1} \int_{-\infty}^{\infty} e^{itx}\, d\Phi^{(1)}(x) \;=\; \ldots \\ &= (-it)^{-j} \int_{-\infty}^{\infty} e^{itx}\, d\Phi^{(j)}(x)\,, \end{aligned}$$

where $\Phi^{(j)}(x) = D^j \Phi(x)$ and D is the differential operator d/dx. Therefore,

$$\int_{-\infty}^{\infty} e^{itx}\, d\{(-D)^j\, \Phi(x)\} = (it)^j\, e^{-t^2/2}\,.$$

Interpreting $r_j(-D)$ as the obvious polynomial in D, we deduce that

$$\int_{-\infty}^{\infty} e^{itx}\, d\{r_j(-D)\, \Phi(x)\} = r_j(it)\, e^{-t^2/2}\,.$$

Therefore, by the uniqueness of Fourier transforms,

$$R_j(x) = r_j(-D)\, \Phi(x)\,.$$

Hermite polynomials

The Hermite polynomials,

$$\begin{aligned}
&\mathrm{He}_0(x) = 1\,, \quad \mathrm{He}_1(x) = x\,,\\
&\mathrm{He}_2(x) = x^2 - 1\,,\\
&\mathrm{He}_3(x) = x\,(x^2 - 3)\,,\\
&\mathrm{He}_4(x) = x^4 - 6\,x^2 + 3\,,\\
&\mathrm{He}_5(x) = x\,(x^4 - 10\,x^2 + 15)\,, \ldots
\end{aligned}$$

are orthogonal with respect the standard Normal density, $\phi = \Phi'$; are normalised so that the coefficient of the term of highest degree is 1; and have the same parity as their index. Note too that He_j is of precise degree j.

Most importantly, from our viewpoint,

$$(-D)^j\,\Phi(x) = -\mathrm{He}_{j-1}(x)\,\phi(x)\,.$$

Formula for R_j

Therefore, if

$$r_j(u) = c_1\, u + \ldots + c_{3j}\, u^{3j}\,,$$

then

$$\begin{aligned} R_j(x) &= r_j(-D)\, \Phi(x) \\ &= -\Big\{c_1\, \mathsf{He}_0(x) + \ldots + c_{3j}\, \mathsf{He}_{3j-1}(u)\Big\} \\ &\qquad \times \phi(x)\,. \end{aligned}$$

It follows that we may write $R_j(x) = P_j(x)\, \phi(x)$, where P_j is a polynomial.

Since r_j is of degree $3j$ and has the same parity as its index; and He_j is of degree j and has the same parity as its index; then P_j is of degree $3j-1$ and has opposite parity to its index. Its coefficients depend on moments of X up to those of order $j+2$.

Formula for R_j (continued)

Examples:

$$\begin{aligned} R_1(x) &= -\tfrac{1}{6}\,\kappa_3\,(x^2-1)\,\phi(x)\,, \\ R_2(x) &= -x\left\{\tfrac{1}{24}\,\kappa_4\,(x^2-3)\right. \\ &\qquad\qquad \left. + \tfrac{1}{72}\,\kappa_3^2\,(x^4-10\,x^2+15)\right\}\phi(x). \end{aligned}$$

Exercise: Derive these formulae. (This is straightforward, given what we have proved already.)

Asymptotic expansions

We have given an heuristic derivation of an expansion of the distribution function of S_n:

$$\begin{aligned} P(S_n \le x) \; &= \; \Phi(x) + n^{-1/2} \, R_1(x) + \dots \\ &\qquad + n^{-j/2} \, R_j(x) + \dots \,, \end{aligned}$$

where

$$R_j(x) = r_j(-D)\, \Phi(x) \,.$$

In order to describe its rigorous form, we must first consider how to interpret the expansion.

The expansion seldom converges as an infinite series. A sufficient condition, due to Cramér, is that $E(e^{X^2/4}) < \infty$, which is rarely true for distributions which are not very closely connected to the Normal distribution.

Asymptotic expansions (continued, 1)

Nevertheless, the expansion does make sense when interpreted as an *asymptotic* series, where the remainder after stopping the expansion after a finite number of terms is of smaller order then the last included term:

$$\begin{aligned} P(S_n \le x) \;&=\; \Phi(x) + n^{-1/2}\, R_1(x) + \ldots \\ &\qquad\qquad + n^{-j/2}\, R_j(x) + o(n^{-j/2})\,. \end{aligned}$$

A sufficient regularity condition for this result is

$$E(|X|^{j+2}) < \infty\,, \qquad \limsup_{|t|\to\infty} |\chi(t)| < 1\,.$$

Rigorous derivation of the expansion under these restrictions was first achieved by Cramér.

When these conditions hold the expansion is valid uniformly in x.

Asymptotic expansions (continued, 2)

Since moments of order $j+2$ appear among the coefficients of the polynomial P_j, and since $R_j = P_j \, \phi$, then the condition $E(|X|^{j+2}) < \infty$ is hard to weaken. It can be relaxed when j is odd, however.

The second condition, $\limsup_{|t|\to\infty} |\chi(t)| < 1$, is called "Cramér's continuity condition." It holds if the distribution function F of X can be written as $F = \pi \, G + (1-\pi) \, H$, where G is the distribution function of a random variable with an absolutely continuous distribution, H is another distribution function, and $0 < \pi \le 1$.

Exercise: Prove that if the distribution F of X is absolutely continuous, i.e. for a density function f,

$$F(x) = \int_{-\infty}^{x} f(u) \, du \,,$$

then Cramér's continuity condition holds in the strong form, $\limsup_{|t|\to\infty} |\chi(t)| = 0$*. Hence, verify the claim made above.*

Asymptotic expansions (continued, 3)

Therefore, Cramér's continuity condition is an assumption about the smoothness of the distribution of X. It fails if the distribution is of lattice type, i.e. if all points x in the support of the distribution of X have the form $x = jh + a$, where $h > 0$ and $-\infty < a < \infty$ are fixed and j is an integer. (If h is as large as possible such that these constraints hold, it is called the "span" of the distribution of X.)

When X has a lattice distribution, with sufficiently many finite moments, an Edgeworth expansion of the distribution of S_n still holds in the form

$$\begin{aligned} P(S_n \leq x) \;&=\; \Phi(x) + n^{-1/2}\,R_1(x) + \ldots \\ &\qquad\qquad + n^{-j/2}\,R_j(x) + o(n^{-j/2})\,, \end{aligned}$$

but the functions R_j have a more complex form. In particular, they are no longer continuous.

Asymptotic expansions (continued, 4)

The “gap” between cases where Cramér’s continuity condition holds, and the case where X has a lattice distribution, is well understood only for $j = 1$. There was shown by Esseen that the expansion

$$P(S_n \leq x) = \Phi(x) + n^{-1/2}\, R_1(x) + o(n^{-1/2})$$

is valid under the sole conditions that the distribution of X is nonlattice and $E(|X|^3) < \infty$.

Asymptotic expansions of densities

Cramér's continuity condition holds in many cases where the distribution of S_n does not have a well-defined density. Therefore, it is unrealistic to expect that an expansion of the distribution of S_n will automatically imply an expansion of its density. However, such an expansion is valid provided S_n has a well-defined density for some n.

There, writing $f_n(x) = (d/dx)\, P(S_n \le x)$, we have:

$$\begin{aligned} f_n(x) &= \phi(x) + n^{-1/2}\, R_1'(x) + \ldots \\ &\qquad + n^{-j/2}\, R_j'(x) + o(n^{-j/2})\,, \end{aligned}$$

provided $E(|X|^{j+2}) < \infty$. The expansion holds uniformly in x.

Asymptotic expansions of densities (continued)

A version of this “local” expansion, as it is called, also holds for lattice distributions, in the form:

$$\begin{aligned} P(S_n = x) &= \phi(x) + n^{-1/2} R_1'(x) + \ldots \\ &\qquad + n^{-j/2} R_j'(x) + o(n^{-j/2}), \end{aligned}$$

uniformly in points x in the support of the distribution.

Curiously, the functions R_n in this expansion are the same ones that appear in the usual, non-lattice expansion of $P(S_n \le x)$.

Exercise: Derive the version of this local lattice expansion in the case where the unstandardised form of X has the Binomial Bi(n, p) *distribution. (Treating the case $j = 1$ is adequate; larger j is similar, but more algebraically complex.) [Hint: Use an expansion related to Stirling’s formula to approximate $\binom{n}{r}$.]*

Expansions in more general cases

The year 2003 was the 75th anniversary of the publication of Cramér's paper, "On the composition of elementary errors," in which he gave the first general, rigorous expansion of the distribution of a sum of independent and identically distributed random variables. The cases of other statistics have been discussed for many years, but it was not until relatively recently, in a pathbreaking paper in 1978 by Bhattacharya and Ghosh, that rigour was provided in a wide range of cases.

Expansions in more general cases (continued, 1)

Bhattacharya and Ghosh dealt with statistics which can be represented as a smooth function, A, of a vector mean, $\bar{X}$; that is, with $A(\bar{X})$ where

$$\begin{aligned} A(x) &= \{g(x) - g(\mu)\}/h(\mu) \quad \text{or} \\ A(x) &= \{g(x) - g(\mu)\}/h(x)\,, \end{aligned}$$

g and h are smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$, $h(\mu) > 0$, and $\bar{X} = n^{-1} \sum_i X_i$ is the mean of the first n of independent and identically distributed random d-vectors $X_1, X_2, \ldots$ with mean μ. (We make these assumptions below.)

Let $t = (t^{(1)}, \ldots, t^{(d)})^{\mathsf{T}}$ denote a d-vector, and let $\chi(t) = E\{\exp(it^{\mathsf{T}}X)\}$ be the characteristic function of the d-vector X, distributed as X_j.

The two different versions of $A(x)$ above allow us to treat "non-Studentised" and "Studentised" cases, respectively.

Expansions in more general cases (continued, 2)

Let $\sigma^2 > 0$ denote the asymptotic variance of $U_n = n^{1/2} A(\bar{X})$, and put $S_n = U_n / \sigma$.

Theorem. (Essentially, Bhattacharya & Ghosh, 1978.) Assume the function A has $j+2$ continuous derivatives in a neighbourhood of μ, and that

$$E(\|X\|^{j+2}) < \infty\,, \qquad \limsup_{\|t\|\to\infty} |\chi(t)| < 1\,.$$

Then,

$$\begin{aligned} P(S_n \le x) \;=\; & \Phi(x) + n^{-1/2}\, R_1(x) + \ldots \\ & \qquad + n^{-j/2}\, R_j(x) + o(n^{-j/2})\,, \end{aligned}$$

uniformly in x, where $R_k(x) = P_k(x)\,\phi(x)$ and P_k is a polynomial of degree $3k-1$, with opposite parity to its index, and with coefficients depending on moments of X up to order $k+2$ and on derivatives of A (evaluated at μ) up to the $(k+2)$nd.

Expansions in more general cases (continued, 3)

Note: x here is a scalar, not a d-vector, and ϕ is the univariate standard Normal density.

For a proof, see:

Bhattacharya, R.N. & Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–451.

The polynomials R_j are identified by developing a Taylor approximation to S_n, of the form

$$S_n = Q_n(\bar{X} - \mu) + O_p(n^{-(j+1)/2}) \,,$$

where Q_n is a polynomial of degree $j+1$. Here we use the fact that:

$$\begin{aligned} A(\bar{X}) &= A(\mu + \bar{X} - \mu) \\ &= A(\mu) + (\bar{X} - \mu)^{\top} \dot{A}(\mu) \\ &\qquad + \tfrac{1}{2}\,(\bar{X} - \mu)^{\top} \ddot{A}(\mu)(\bar{X} - \mu) + \ldots \end{aligned}$$

Expansions in more general cases (continued, 4)

Since Q_n is a polynomial and $\bar{X}$ is a sample mean then the cumulants of the distribution of $Q_n(\bar{X}-\mu)$ can be written down fairly easily, and hence a formal expansion of the distribution of $Q_n(\bar{X}-\mu)$ can be developed, up to j terms:

$$\begin{aligned} P\{Q_n(\bar{X}-\mu) \le x\} \\ &= \Phi(x) + n^{-1/2}\, R_1(x) + \ldots \\ &\quad + n^{-j/2}\, R_j(x) + o(n^{-j/2})\,, \end{aligned}$$

The functions R_j appearing here are exactly those appearing in the Taylor expansion of the distribution of S_n.

Studentised and non-Studentised cases

Expansions in Studentised and non-Studentised cases have different polynomials. For example, in the case of the Studentised mean,

$$\begin{aligned} P_1(x) &= \tfrac{1}{6}\,\kappa_3\,(2x^2+1)\,, \\ P_2(x) &= x\left\{\tfrac{1}{12}\,\kappa_4\,(x^2-3)\right. \\ &\qquad -\tfrac{1}{18}\,\kappa_3^2\,(x^4+2\,x^2-3) \\ &\qquad \left.-\tfrac{1}{4}\,(x^2+3)\right\}. \end{aligned}$$

We know already that in the non-Studentised case,

$$\begin{aligned} P_1(x) &= -\tfrac{1}{6}\,\kappa_3\,(x^2-1)\,, \\ P_2(x) &= -x\left\{\tfrac{1}{24}\,\kappa_4\,(x^2-3)\right. \\ &\qquad \left.+\tfrac{1}{72}\,\kappa_3^2\,(x^4-10\,x^2+15)\right\}. \end{aligned}$$

Cornish-Fisher expansions

We have shown how to develop *Edgeworth expansions* of the distribution of a statistic S_n:

$$\begin{aligned} P(S_n \le x) &= \Phi(x) + n^{-1/2} R_1(x) + \ldots \\ &\qquad + n^{-j/2} R_j(x) + o(n^{-j/2})\,. \end{aligned}$$

This is an expansion of a probability for a given value of a quantile, x. Defining ξ_α to be the solution of

$$P(S_n \le \xi_\alpha) = \alpha\,,$$

for a given, fixed value of $\alpha \in (0,1)$, we may "invert" the expansion to express ξ_α as a series expansion:

$$\begin{aligned} \xi_\alpha &= z_\alpha + n^{-1/2} P_1^{\mathrm{cf}}(z_\alpha) + \ldots \\ &\qquad + n^{-j/2} P_j^{\mathrm{cf}}(z_\alpha) + o(n^{-j/2})\,, \end{aligned}$$

where $z_\alpha = \Phi^{-1}(\alpha)$ denotes the α-level quantile of the standard Normal distribution and $P_1^{\mathrm{cf}}, P_2^{\mathrm{cf}}$ etc are polynomials.

Cornish-Fisher expansions (continued)

Noting that $R_j = P_j \, \Phi$ for polynomials P_j, it may be proved that $P_1^{\text{cf}} = -P_1$,

$$P_2^{\text{cf}}(x) = P_1(x)\, P_1'(x) - \tfrac{1}{2}\, x\, P_1(x)^2 - P_2(x)\,,$$

etc.

Exercise: Prove these formulae.

METHODOLOGY AND THEORY FOR THE BOOTSTRAP
(Fourth set of two lectures)

Main topic of these lectures: Theoretical properties of bootstrap confidence intervals

Studentised and non-Studentised estimators

Let $\widehat{\theta} = \theta(\widehat{F})$ denote the bootstrap estimator of a statistic $\theta = \theta(F)$, computed from a dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$. (Here, F denotes the distribution function of the data X_i.)

Write $\sigma^2 = \sigma^2(F)$ for the asymptotic variance of

$$S = n^{1/2}\,(\widehat{\theta} - \theta)\,,$$

which we assume has a limiting Normal N$(0, \sigma^2)$ distribution.

Studentised and non-Studentised estimators (continued)

Let $\widehat{\sigma}^2 = \sigma^2(\widehat{F})$ denote the bootstrap estimator of σ^2. The "Studentised" form of S is

$$T = S/\widehat{\sigma} = n^{1/2}\,(\widehat{\theta} - \theta)/\widehat{\sigma}\,,$$

which has a limiting Normal $\mathrm{N}(0,1)$ distribution. Therefore,

$$\begin{aligned} P(S \le \sigma x) &= \Phi(x) + o(1)\,, \\ P(T \le x) &= \Phi(x) + o(1)\,. \end{aligned}$$

We say that T is (asymptotically) pivotal, because its limiting distribution does not depend on unknowns.

Edgeworth expansions of distributions of S and T

We know from previous lectures that, in a wide range of settings studied by Bhattacharya and Ghosh, the distributions of S and T admit Edgeworth expansions:

$$\begin{aligned} P(S \le \sigma x) &= \Phi(x) + n^{-1/2} P_1(x)\,\phi(x) \\ &\qquad + n^{-1} P_2(x)\,\phi(x) + \ldots\,, \\ P(T \le x) &= \Phi(x) + n^{-1/2} Q_1(x)\,\phi(x) \\ &\qquad + n^{-1} Q_2(x)\,\phi(x) + \ldots\,, \end{aligned}$$

where P_j and Q_j are polynomials of degree $3j-1$, of opposite parity to their indices.

Bootstrap Edgeworth expansions

Let $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denote a resample drawn by sampling randomly, with replacement, from $\mathcal{X}$; and let $\widehat{\theta}^*$ and $\widehat{\sigma}^*$ be the same functions of the bootstrap data $\mathcal{X}^*$ as $\widehat{\theta}$ and $\widehat{\sigma}$ were of the real data $\mathcal{X}$. Put

$$\begin{aligned} S^* &= n^{1/2}\,(\widehat{\theta}^* - \widehat{\theta})\,, \\ T^* &= n^{1/2}\,(\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*\,, \end{aligned}$$

denoting the bootstrap versions of S and T. The bootstrap distributions of these quantities are their distributions conditional on $\mathcal{X}$, and they admit analogous Edgeworth expansions:

$$\begin{aligned} P(S^* \le \widehat{\sigma} x \mid \mathcal{X}) &= \Phi(x) + n^{-1/2}\,\widehat{P}_1(x)\,\phi(x) \\ &\qquad + n^{-1}\,\widehat{P}_2(x)\,\phi(x) + \ldots\,, \\ P(T^* \le x \mid \mathcal{X}) &= \Phi(x) + n^{-1/2}\,\widehat{Q}_1(x)\,\phi(x) \\ &\qquad + n^{-1}\,\widehat{Q}_2(x)\,\phi(x) + \ldots\,. \end{aligned}$$

In these formulae, $\widehat{P}_j$ and $\widehat{Q}_j$ are the versions of P_j and Q_j in which unknown quantities are replaced by their bootstrap estimators.

Bootstrap Edgeworth expansions (continued, 1)

For example, recall that when θ and σ^2 denote the population mean and variance,

$$\begin{aligned} P_1(x) &= -\tfrac{1}{6}\,\kappa_3\,(x^2-1)\,, \\ P_2(x) &= -x\left\{\tfrac{1}{24}\,\kappa_4\,(x^2-3)\right. \\ &\qquad \left. + \tfrac{1}{72}\,\kappa_3^2\,(x^4-10\,x^2+15)\right\}, \\ Q_1(x) &= \tfrac{1}{6}\,\kappa_3\,(2x^2+1)\,, \\ Q_2(x) &= x\left\{\tfrac{1}{12}\,\kappa_4\,(x^2-3)\right. \\ &\qquad - \tfrac{1}{18}\,\kappa_3^2\,(x^4+2\,x^2-3) \\ &\qquad \left. -\tfrac{1}{4}\,(x^2+3)\right\}. \end{aligned}$$

Replace $\kappa_3 = \sigma^{-3}\,E(X-EX)^3$ and $\kappa_4 = \sigma^{-4}\,E(X-EX)^4 - 3$ by their bootstrap estimators,

$$\begin{aligned} \widehat{\kappa}_3 &= \widehat{\sigma}^{-3}\,n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^3\,, \\ \widehat{\kappa}_4 &= \widehat{\sigma}^{-4}\,n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^4 - 3\,, \end{aligned}$$

to get $\widehat{P}_1, \widehat{P}_2, \widehat{Q}_1, \widehat{Q}_2$.

Skewness and kurtosis

Note that we call κ_3 "skewness," and κ_4 "kurtosis." Therefore, the adjustment of order $n^{-1/2}$ that an Edgeworth expansion applies to the standard Normal approximation is a correction arising from skewness, i.e. from asymmetry. If skewness was zero (i.e. if $\kappa_3 = 0$), and in particular if the sampled distribution was symmetric, then the term of order $n^{-1/2}$ would vanish, and the first nonzero term appearing in the Edgeworth expansion would be of size n^{-1}.

Likewise, the term of size n^{-1} in the expansion is a second-order correction for skewness, and a first-order correction for kurtosis, or tail weight. (Kurtosis describes the difference between the weight of the tails of the sampled distribution and that of the Normal distribution with the same mean and variance. If $\kappa_4 > 0$ then the tails of the sampled distribution tend to be heavier, and if $\kappa_4 < 0$ they tend to be lighter.)

Skewness and kurtosis (continued)

We can make the same interpretation for Edgeworth expansions in general cases, not just the case of the sample mean. In general, the term of size $n^{-1/2}$ in an Edgeworth expansion provides a first-order correction for asymmetry of the sampled distribution. The term of size n^{-1} provides a first-order correction for tailweight (relative to that of the Normal distribution) and a second-order correction for asymmetry.

Note particularly that, to first order, the bootstrap correctly captures the effects of asymmetry. In particular, since $\widehat{\kappa}_3 = \kappa_3 + O_p(n^{-1/2})$ then

$$\begin{aligned} \widehat{Q}_1(x) &= \tfrac{1}{6}\,\widehat{\kappa}_3\,(2x^2+1) \\ &= \tfrac{1}{6}\,\kappa_3\,(2x^2+1) + O_p(n^{-1/2}) \\ &= Q_1(x) + O_p(n^{-1/2})\,, \end{aligned}$$

Similarly, $\widehat{P}_j = P_j + O_p(n^{-1/2})$ and $\widehat{Q}_j = Q_j + O_p(n^{-1/2})$ for each j.

Accuracy of bootstrap approximations

It follows that

$$
\begin{aligned}
P(S^* \le \hat{\sigma} x \mid \mathcal{X}) &= \Phi(x) + n^{-1/2}\, \widehat{P}_1(x)\, \phi(x) \\
&\qquad + O_p(n^{-1}) \\
&= \Phi(x) + n^{-1/2}\, P_1(x)\, \phi(x) \\
&\qquad + O_p(n^{-1}) \\
&= P(S \le \sigma x) + O_p(n^{-1})\,, \\
P(T^* \le x \mid \mathcal{X}) &= \Phi(x) + n^{-1/2}\, \widehat{Q}_1(x)\, \phi(x) \\
&\qquad + O_p(n^{-1})\,, \\
&= \Phi(x) + n^{-1/2}\, Q_1(x)\, \phi(x) \\
&\qquad + O_p(n^{-1}) \\
&= P(T \le x) + O_p(n^{-1})\,.
\end{aligned}
$$

That is, the bootstrap distributions of $S^*/\hat{\sigma}$ and T^* approximate the true distributions of S/σ and T, respectively, to orders n^{-1}:

$$
\begin{aligned}
P(S^* \le \hat{\sigma} x \mid \mathcal{X}) &= P(S \le \sigma x) + O_p(n^{-1})\,, \\
P(T^* \le x \mid \mathcal{X}) &= P(T \le x) + O_p(n^{-1})\,.
\end{aligned}
$$

Compare this with the Normal approximation, where accuracy is generally only $O(n^{-1/2})$.

These orders of approximation are valid uniformly in x.

Accuracy of bootstrap approximations (continued)

These results underpin the performance of bootstrap methods in distribution approximation, and show its advantages over conventional Normal approximations.

Note, however, that the bootstrap distribution of $S^*/\hat{\sigma}$ approximates the true distributions of S/σ, *not* the true distribution of $S/\hat{\sigma} = T$. Therefore, in order to effectively use approximations based on S^* we usually have to know the value of σ; but generally we do not.

Therefore, we do not necessarily get such good performance when using the standard "percentile bootstrap," which refers to methods based on S^*, rather than the "percentile-t bootstrap," referring to methods based on T^*.

Cornish-Fisher expansions

Let $\xi_\alpha, \eta_\alpha, \widehat{\xi}_\alpha, \widehat{\eta}_\alpha$ denote α-level quantiles of the distributions of $S/\sigma, T$ and the bootstrap distributions of $S^*/\widehat{\sigma}, T^*$, respectively:

$$\begin{aligned} P(S/\sigma \le \xi_\alpha) = P(T \le \eta_\alpha) = P(S^*/\widehat{\sigma} \le \widehat{\xi}_\alpha \mid \mathcal{X}) \\ = \ P(T^* \le \widehat{\eta}_\alpha \mid \mathcal{X}) = \alpha\,. \end{aligned}$$

Cornish-Fisher expansions of quantiles, in both their conventional and bootstrap forms, are:

$$\begin{aligned} \xi_\alpha &= z_\alpha + n^{-1/2}\, P_1^{\text{cf}}(z_\alpha) + n^{-1}\, P_2^{\text{cf}}(z_\alpha) + \ldots\,, \\ \eta_\alpha &= z_\alpha + n^{-1/2}\, Q_1^{\text{cf}}(z_\alpha) + n^{-1}\, Q_2^{\text{cf}}(z_\alpha) + \ldots\,, \\ \widehat{\xi}_\alpha &= z_\alpha + n^{-1/2}\, \widehat{P}_1^{\text{cf}}(z_\alpha) + n^{-1}\, \widehat{P}_2^{\text{cf}}(z_\alpha) + \ldots\,, \\ \widehat{\eta}_\alpha &= z_\alpha + n^{-1/2}\, \widehat{Q}_1^{\text{cf}}(z_\alpha) + n^{-1}\, \widehat{Q}_2^{\text{cf}}(z_\alpha) + \ldots\,, \end{aligned}$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the standard Normal α-level critical point.

Cornish-Fisher expansions (continued, 1)

Recall that $P_1^{\mathsf{cf}} = -P_1$, $Q_1^{\mathsf{cf}} = -Q_1$,

$$\begin{aligned} P_2^{\mathsf{cf}}(x) &= P_1(x)\,P_1'(x) - \tfrac{1}{2}\,x\,P_1(x)^2 - P_2(x)\,, \\ Q_2^{\mathsf{cf}}(x) &= Q_1(x)\,Q_1'(x) - \tfrac{1}{2}\,x\,Q_1(x)^2 - Q_2(x)\,, \end{aligned}$$

etc. Of course, the bootstrap analogues of these formulae hold too: $\widehat{P}_1^{\mathsf{cf}} = -\widehat{P}_1$, $\widehat{Q}_1^{\mathsf{cf}} = -\widehat{Q}_1$,

$$\begin{aligned} \widehat{P}_2^{\mathsf{cf}}(x) &= \widehat{P}_1(x)\,\widehat{P}_1'(x) - \tfrac{1}{2}\,x\,\widehat{P}_1(x)^2 - \widehat{P}_2(x)\,, \\ \widehat{Q}_2^{\mathsf{cf}}(x) &= \widehat{Q}_1(x)\,\widehat{Q}_1'(x) - \tfrac{1}{2}\,x\,\widehat{Q}_1(x)^2 - \widehat{Q}_2(x)\,, \end{aligned}$$

etc.

Cornish-Fisher expansions (continued, 2)

Therefore, since $\widehat{P}_j = P_j + O_p(n^{-1/2})$ and $\widehat{Q}_j = Q_j + O_p(n^{-1/2})$, it is generally true that

$$\begin{aligned} \widehat{P}_j^{\text{cf}}(x) &= P_j^{\text{cf}}(x) + O_p(n^{-1/2})\,, \\ \widehat{Q}_j^{\text{cf}}(x) &= Q_j^{\text{cf}}(x) + O_p(n^{-1/2})\,, \end{aligned}$$

and hence that

$$\begin{aligned} \hat{\xi}_\alpha &= z_\alpha + n^{-1/2}\,\widehat{P}_1^{\text{cf}}(z_\alpha) + n^{-1}\,\widehat{P}_2^{\text{cf}}(z_\alpha) + \ldots\,, \\ &= \xi_\alpha + O_p(n^{-1})\,, \\ \hat{\eta}_\alpha &= z_\alpha + n^{-1/2}\,\widehat{Q}_1^{\text{cf}}(z_\alpha) + n^{-1}\,\widehat{Q}_2^{\text{cf}}(z_\alpha) + \ldots\,, \\ &= \eta_\alpha + O_p(n^{-1})\,, \end{aligned}$$

(These orders of approximation are valid uniformly in x on compact intervals.)

Again the order of accuracy of the bootstrap approximation is n^{-1}, bettering the order, $n^{-1/2}$, of the conventional Normal approximation. But the same caveat applies: in order for approximations based on the percentile bootstrap, i.e. involving S^*, to be effective, we need to know σ.

Bootstrap confidence intervals

We shall work initially only with one-sided confidence intervals, putting them together later to get two-sided intervals.

The intervals

$$\begin{aligned} I_1 &= (-\infty, \widehat{\theta} - n^{-1/2}\,\sigma\,\xi_{1-\alpha})\,, \\ J_1 &= (-\infty, \widehat{\theta} - n^{-1/2}\,\widehat{\sigma}\,\eta_{1-\alpha}) \end{aligned}$$

cover θ with probability exactly α, but are in general not computable, since we do not know either $\xi_{1-\alpha}$ or $\eta_{1-\alpha}$. On the other hand, their bootstrap counterparts

$$\begin{aligned} \widehat{I}_{11} &= (-\infty, \widehat{\theta} - n^{-1/2}\,\sigma\,\widehat{\xi}_{1-\alpha})\,, \\ \widehat{I}_{12} &= (-\infty, \widehat{\theta} - n^{-1/2}\,\widehat{\sigma}\,\widehat{\xi}_{1-\alpha})\,, \\ \widehat{J}_1 &= (-\infty, \widehat{\theta} - n^{-1/2}\,\widehat{\sigma}\,\widehat{\eta}_{1-\alpha}) \end{aligned}$$

are readily computed from data, but their coverage probabilities are not known exactly. They are respectively called "percentile" and "percentile-t" bootstrap confidence intervals for θ.

Bootstrap confidence intervals (continued)

The "other" percentile confidence intervals for θ are

$$\begin{aligned} K_1 &= (-\infty, \hat{\theta} + n^{-1/2}\,\sigma\,\xi_\alpha)\,, \\ \widehat{K}_1 &= (-\infty, \hat{\theta} + n^{-1/2}\,\hat{\sigma}\,\hat{\xi}_\alpha)\,. \end{aligned}$$

Neither, in general, has exact coverage. The interval $\widehat{K}_1$ is the type of bootstrap confidence interval we introduced early in this series of lectures.

We expect the coverage probabilities of $\hat{I}_1$, $\hat{J}_1$, K_1 and $\widehat{K}_1$ to converge to α as $n \to \infty$. However, the convergence rate is generally only $n^{-1/2}$. The exceptions are I_{11} and $\hat{J}_1$, for which coverage error equals $\alpha + O(n^{-1})$.

The interval I_{11} is generally not particularly useful, since we need to know σ in order to use it effectively. Therefore, of the intervals we have considered, only J_1 is both useful and has good coverage accuracy.

Advantages of using a pivotal statistic

In summary: Unless the asymptotic variance σ^2 is known, one-sided bootstrap confidence intervals based on the pivotal statistic T generally have a higher order of coverage accuracy than intervals based on the non-pivotal statistic S.

Intuitively, this is because (in the case of intervals based on S) the bootstrap spends the majority of its effort implicitly computing a correction for scale. It does not provide an effective correction for skewness; and, as we have seen, the main term describing the departure of the distribution of a statistic from Normality is due to skewness.

On the other hand, when the bootstrap is applied to a a pivotal statistic such as T, which is already corrected for scale, it devotes itself to correcting for skewness, and therefore adjusts for the major part of the error in a Normal approximation.

Derivation of these properties

We begin with the case of the confidence interval

$$\hat{J}_1 = (-\infty, \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{1-\alpha})\,,$$

our aim being to show that

$$P(\theta \in \hat{J}_1) = \alpha + O(n^{-1})\,.$$

Recall that $\hat{\eta}_{1-\alpha}$ is defined by

$$P(T^* \le \hat{\eta}_{1-\alpha} \mid \mathcal{X}) = 1 - \alpha\,,$$

and that, by a Cornish-Fisher expansion,

$$\begin{aligned} \hat{\eta}_{1-\alpha} &= z_{1-\alpha} + n^{-1/2}\,\hat{Q}_1^{\text{cf}}(z_{1-\alpha}) + O_p(n^{-1}) \\ &= z_{1-\alpha} + n^{-1/2}\,Q_1^{\text{cf}}(z_{1-\alpha}) + O_p(n^{-1}) \\ &= \eta_{1-\alpha} + O_p(n^{-1})\,, \end{aligned}$$

where $\eta_{1-\alpha}$ is defined by $P(T \le \eta_{1-\alpha}) = 1-\alpha$. Therefore,

$$\begin{aligned} P(\theta \in \hat{J}_1) &= P\{n^{1/2}(\hat{\theta} - \theta)/\hat{\sigma} > \hat{\eta}_{1-\alpha}\} \\ &= P\left\{n^{1/2}(\hat{\theta} - \theta)/\hat{\sigma} > \eta_{1-\alpha} + O_p(n^{-1})\right\}. \end{aligned}$$

Derivation (continued, 1)

$$P(\theta \in \hat{J}_1) = P\left\{n^{1/2}(\hat{\theta} - \theta)/\hat{\sigma} > \eta_{1-\alpha} + O_p(n^{-1})\right\}$$
$$= P\{T > \eta_{1-\alpha} + O_p(n^{-1})\}.$$

If the $O_p(n^{-1})$ term, on the right-hand side, were a constant rather than a random variable, it would be straightforward to show, using the property

$$\begin{aligned} P(T \le x) &= \Phi(x) + n^{-1/2} Q_1(x)\,\phi(x) \\ &\qquad + O(n^{-1})\,, \end{aligned}$$

on taking $x = \eta_{1-\alpha} + O_p(n^{-1})$, that

$$\begin{aligned} &P\{T \le \eta_{1-\alpha} + O_p(n^{-1})\} \\ &\quad = \Phi\{\eta_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad\quad + n^{-1/2} Q_1\{\eta_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad\qquad \times \phi\{\eta_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad\quad + O(n^{-1}) \\ &\quad = \Phi(\eta_{1-\alpha}) + n^{-1/2} Q_1(\eta_{1-\alpha})\,\phi(\eta_{1-\alpha}) \\ &\qquad\qquad + O(n^{-1}) \\ &\quad = P(T \le \eta_{1-\alpha}) + O(n^{-1})\,. \end{aligned}$$

(These steps need only Taylor expansion.) Therefore,

$$1 - P(\theta \in \hat{J}_1) = P(T \le \eta_{1-\alpha}) + O(n^{-1})\,.$$

Derivation (continued, 2)

This step can be justified using a longer argument, which we shall not give here. Therefore,

$$\begin{aligned} 1 - P(\theta \in \hat{J}_1) &= P\{T \le \eta_{1-\alpha} + O_p(n^{-1})\} \\ &= P(T \le \eta_{1-\alpha}) + O(n^{-1}) \\ &= 1 - \alpha + O(n^{-1})\,, \end{aligned}$$

the last line following from the definition of η_α. That is,

$$P(\theta \in \hat{J}_1) = \alpha + O(n^{-1})\,,$$

which proves that the coverage error of the confidence interval J_1 equals $O(n^{-1})$.

Comparison with Normal-approximation interval

A similar argument shows that coverage error for the corresponding confidence interval based on a Normal approximation, i.e.

$$\begin{aligned} N_1 &= (-\infty, \widehat{\theta} - n^{-1/2}\, \widehat{\sigma}\, z_{1-\alpha}) \\ &= (-\infty, \widehat{\theta} + n^{-1/2}\, \widehat{\sigma}\, z_{\alpha})\,, \end{aligned}$$

equals only $O(n^{-1/2})$.

Exercise. Prove that

$$P(\theta \in N_1) = \alpha - n^{-1/2}\, Q_1(z_\alpha)\, \phi(z_\alpha) + O(n^{-1})\,.$$

Therefore, unless the effect of skewness is vanishingly small (i.e. the polynomial Q_1 vanishes), coverage error of the classical Normal approximation interval, N_1, is an order of magnitude greater than for the percentile-t interval $\widehat{J}_1$.

Derivation (continued, 3)

Similarly it can be proved that coverage error of the interval

$$\hat{I}_{11} = (-\infty, \hat{\theta} - n^{-1/2}\,\sigma\,\hat{\xi}_{1-\alpha})$$

also equals $\alpha + O(n^{-1})$:

$$P(\theta \in \hat{I}_{11}) = \alpha + O(n^{-1})\,.$$

However, this result fails for the more practical interval

$$\hat{I}_{12} = (-\infty, \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\xi}_{1-\alpha})\,,$$

as we now show. The argument will highlight differences between Edgeworth expansions in Studentised and non-Studentised (i.e. pivotal and non-pivotal) cases.

Derivation (continued, 4)

Recall that $\widehat{\xi}_{1-\alpha}$ is defined by

$$P(S^* \le \widehat{\sigma}\, \widehat{\xi}_{1-\alpha} \mid \mathcal{X}) = 1 - \alpha \,,$$

and that, by a Cornish-Fisher expansion,

$$\begin{aligned} \widehat{\xi}_{1-\alpha} &= z_{1-\alpha} + n^{-1/2}\, \widehat{P}_1^{\text{cf}}(z_{1-\alpha}) + O_p(n^{-1}) \\ &= z_{1-\alpha} + n^{-1/2}\, P_1^{\text{cf}}(z_{1-\alpha}) + O_p(n^{-1}) \\ &= \xi_{1-\alpha} + O_p(n^{-1}) \,, \end{aligned}$$

where $\xi_{1-\alpha}$ is defined by $P(S \le \sigma\, \xi_{1-\alpha}) = 1 - \alpha$. Therefore,

$$\begin{aligned} P(\theta \in \widehat{I}_{12}) &= P\{n^{1/2}(\widehat{\theta} - \theta)/\widehat{\sigma} > \widehat{\xi}_{1-\alpha}\} \\ &= P\left\{n^{1/2}(\widehat{\theta} - \theta)/\widehat{\sigma} > \xi_{1-\alpha} + O_p(n^{-1})\right\} \\ &= P\{T > \xi_{1-\alpha} + O_p(n^{-1})\} \,. \end{aligned}$$

Derivation (continued, 5)

By Taylor expansion,

$$\begin{aligned} P\{T \le \xi_{1-\alpha} + O_p(n^{-1})\} &= \Phi\{\xi_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad + n^{-1/2}\, Q_1\{\xi_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad\qquad \times \phi\{\xi_{1-\alpha} + O_p(n^{-1})\} \\ &\qquad + O(n^{-1}) \\ &= \Phi(\xi_{1-\alpha}) + n^{-1/2}\, Q_1(\xi_{1-\alpha})\, \phi(\xi_{1-\alpha}) \\ &\qquad + O(n^{-1}) \\ &= P(T \le \xi_{1-\alpha}) + O(n^{-1}) \\ &= P(T \le \eta_{1-\alpha}) + (\xi_{1-\alpha} - \eta_{1-\alpha})\, \phi(\eta_{1-\alpha}) \\ &\qquad + O(n^{-1})\,. \end{aligned}$$

Note too that

$$\begin{aligned} \xi_{1-\alpha} - \eta_{1-\alpha} &= n^{-1/2}\, \{P_1^{\mathrm{cf}}(z_{1-\alpha}) - Q_1^{\mathrm{cf}}(z_{1-\alpha})\} + O(n^{-1}) \\ &= n^{-1/2}\, \{Q_1(z_\alpha) - P_1(z_\alpha)\} + O(n^{-1})\,. \end{aligned}$$

(Recall that $P_1^{\mathrm{cf}} = -P_1$, $Q_1^{\mathrm{cf}} = -Q_1$, and P_1 and Q_1 are both even polynomials.)

Derivation (continued, 6)

Hence,

$$\begin{aligned} P\{T \le \xi_{1-\alpha} + O_p(n^{-1})\} &= P(T \le \eta_{1-\alpha}) \\ &\qquad + n^{-1/2}\{Q_1(z_\alpha) - P_1(z_\alpha)\}\,\phi(z_\alpha) \\ &\qquad + O(n^{-1}) \\ &= 1 - \alpha + n^{-1/2}\{Q_1(z_\alpha) - P_1(z_\alpha)\}\,\phi(z_\alpha) \\ &\qquad + O(n^{-1})\,. \end{aligned}$$

Therefore,

$$\begin{aligned} P(\theta \in \hat{I}_{12}) &= P\{T > \xi_{1-\alpha} + O_p(n^{-1})\} \\ &= \alpha + n^{-1/2}\{P_1(z_\alpha) - Q_1(z_\alpha)\}\,\phi(z_\alpha) \\ &\qquad + O(n^{-1})\,. \end{aligned}$$

It follows that the confidence interval $\hat{I}_{12}$ has coverage error of order n^{-1} if and only if

$$P_1(z_\alpha) - Q_1(z_\alpha) = 0\,;$$

that is, if and only if the skewness terms, in Edgeworth expansions of the distributions of Studentised and non-Studentised forms of the statistic, are identical when evaluated at z_α.

Two-sided confidence intervals

Equal-tailed, two-sided confidence intervals are usually obtained by combining two one-sided intervals. For example, if we are using the interval

$$\hat{J}_1 = \hat{J}_1(\alpha) = (-\infty, \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{1-\alpha})\,,$$

for which the nominal coverage is α, we would generally construct from it the two-sided interval

$$\begin{aligned} \hat{J}_2 &= \hat{J}_1\left\{\tfrac{1}{2}(1+\alpha)\right\} \setminus \hat{J}_1\left\{\tfrac{1}{2}(1-\alpha)\right\} \\ &= \Big[\hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{(1+\alpha)/2}\,, \\ &\qquad\qquad \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{(1-\alpha)/2}\Big)\,. \end{aligned}$$

The actual coverage of $\hat{J}_2$ equals $\alpha + O(n^{-1})$, and can be derived as follows.

Two-sided confidence intervals (continued)

$$\begin{aligned}
P(\theta \in \hat{J}_2) &= P\left[\theta \in \hat{J}_1\left\{\tfrac{1}{2}(1+\alpha)\right\}\right] \\
&\qquad - P\left[\theta \in \hat{J}_1\left\{\tfrac{1}{2}(1-\alpha)\right\}\right] \\
&= \left\{\tfrac{1}{2}(1+\alpha) + O(n^{-1})\right\} \\
&\qquad - \left\{\tfrac{1}{2}(1+\alpha) + O(n^{-1})\right\} \\
&= \alpha + O(n^{-1})\,.
\end{aligned}$$

However, the two-sided version of the percentile confidence interval $\hat{I}_{12}$, which has coverage error only $O(n^{-1/2})$ in its one-sided form, nevertheless has coverage error $O(n^{-1})$. This is a consequence of parity properties of polynomials appearing in Edgeworth expansions.

METHODOLOGY AND THEORY FOR THE BOOTSTRAP (Fifth set of two lectures)

Main topic of these lectures: Completion of work on confidence intervals, and survey of miscellaneous topics

Revision of confidence intervals

Recall that $\widehat{\eta}_\alpha$ is the α-level quantile of the bootstrap distribution of $T^* = n^{1/2}(\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*$:

$$P(T^* \leq \widehat{\eta}_\alpha \mid \mathcal{X}) = \alpha \,.$$

A one-sided percentile-t confidence interval for an unknown parameter θ, based on the bootstrap estimator $\widehat{\theta}$ and having nominal coverage α, is therefore

$$\hat{J}_1 = \hat{J}_1(\alpha) = (-\infty, \widehat{\theta} - n^{-1/2}\,\widehat{\sigma}\,\widehat{\eta}_{1-\alpha}) \,.$$

Revision (continued)

It has coverage error $O(n^{-1})$:

$$P\{\theta \in \hat{J}_1(\alpha)\} = \alpha + O(n^{-1})\,.$$

A conventional two-sided interval, for which the nominal coverage is also α, is obtained from two one-sided intervals:

$$\begin{aligned}\hat{J}_2(\alpha) &= \hat{J}_1\left\{\tfrac{1}{2}(1+\alpha)\right\} \setminus \hat{J}_1\left\{\tfrac{1}{2}(1-\alpha)\right\} \\ &= \Big[\hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{(1+\alpha)/2}\,, \\ &\qquad\qquad \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,\hat{\eta}_{(1-\alpha)/2}\Big)\,.\end{aligned}$$

Unsurprisingly, the actual coverage of $\hat{J}_2$ also equals $\alpha + O(n^{-1})$:

$$P\{\theta \in \hat{J}_2(\alpha)\} = \alpha + O(n^{-1})\,.$$

Percentile method intervals

Interestingly, however, this result extends to the case of two-sided percentile intervals, the one-sided versions of which have coverage accuracy only $O(n^{-1/2})$. Recall that one form of percentile-method confidence interval for θ is

$$\hat{I}_{12}(\alpha) = (-\infty, \hat{\theta} - n^{-1/2}\, \hat{\sigma}\, \hat{\xi}_{1-\alpha})\,,$$

where $\hat{\xi}_\alpha$ is the α-level critical point of the bootstrap distribution of $S^*/\hat{\sigma}$:

$$P(S^*/\hat{\sigma} \le \hat{\xi}_\alpha \mid \mathcal{X}) = \alpha\,.$$

The corresponding two-sided interval is

$$\begin{aligned}
\hat{I}_{22}(\alpha) &= \hat{I}_{12}\left\{\tfrac{1}{2}(1+\alpha)\right\} \setminus \hat{I}_{12}\left\{\tfrac{1}{2}(1-\alpha)\right\} \\
&= \Big[\hat{\theta} - n^{-1/2}\, \hat{\sigma}\, \hat{\xi}_{(1+\alpha)/2}\,, \\
&\qquad\qquad \hat{\theta} - n^{-1/2}\, \hat{\sigma}\, \hat{\xi}_{(1-\alpha)/2}\Big)\,.
\end{aligned}$$

Coverage of two-sided percentile intervals

To calculate the coverage of $\hat{I}_{22}(\alpha)$, recall that

$$\begin{aligned} &P\{\theta \in \hat{I}_{12}(\alpha)\} \\ &\quad = \alpha + n^{-1/2}\,\{P_1(z_\alpha) - Q_1(z_\alpha)\}\,\phi(z_\alpha) \\ &\qquad + O(n^{-1})\,. \end{aligned}$$

Since P_1 and Q_1 are even polynomials, and $z_{(1+\alpha)/2} = -z_{(1-\alpha)/2}$, then

$$\begin{aligned} &P_1(z_{(1+\alpha)/2}) - Q_1(z_{(1+\alpha)/2}) \\ &\qquad = P_1(z_{(1-\alpha)/2}) - Q_1(z_{(1-\alpha)/2}) \end{aligned}$$

Therefore,

$$\begin{aligned} &P\{\theta \in \hat{I}_{22}(\alpha)\} \\ &\quad = P\left[\theta \in \hat{I}_{12}\left\{\tfrac{1}{2}\,(1+\alpha)\right\}\right] \\ &\qquad -P\left[\theta \in \hat{I}_{12}\left\{\tfrac{1}{2}\,(1-\alpha)\right\}\right] \\ &\quad = \tfrac{1}{2}\,(1+\alpha) - \tfrac{1}{2}\,(1-\alpha) + O(n^{-1}) \\ &\quad = \alpha + O(n^{-1})\,. \end{aligned}$$

Coverage of two-sided percentile intervals (continued)

Therefore, owing to the parity properties of polynomials in Edgeworth expansions, this two-sided percentile confidence interval has coverage error $O(n^{-1})$. The same result holds true for the "other" type of percentile confidence interval, of which the one-sided form is

$$\widehat{K}_1(\alpha) = (-\infty, \hat\theta + n^{-1/2}\,\hat\sigma\,\hat\xi_\alpha)\,.$$

Its one- and two-sided forms have coverage

$$\begin{aligned} P\{\theta \in \widehat{K}_1(\alpha)\} &= \alpha + O(n^{-1/2})\,, \\ P\{\theta \in \widehat{K}_2(\alpha)\} &= \alpha + O(n^{-1})\,. \end{aligned}$$

Exercise: (1) Derive the latter property.

(2) Show that, when computing percentile confidence intervals, as distinct from percentile-t intervals, we do not actually need the value of $\hat\sigma$. (It has been included for didactic reasons, to clarify our presentation of theory, but it cancels in numerical calculations.)

Discussion

Therefore, the arguments in favour of percentile-t methods are less powerful when applied to two-sided confidence intervals. However, the asymmetry of percentile intervals will usually not accurately reflect that of the statistic $\widehat{\theta}$, and in this sense they are less appropriate.

This is especially true in the case of the intervals $\widehat{K}$ ("the other percentile method"). There, when $\widehat{\theta}$ has a markedly asymmetric distribution, the lengths of the two sides of a two-sided interval based on $\widehat{K}_1$ will reflect the exact opposite of the tailweights.

Other bootstrap confidence intervals

It is possible to correct bootstrap confidence intervals for skewness without Studentising. The best-known examples of this type are the "accelerated bias corrected" intervals proposed by Bradley Efron, based on explicit corrections for skewness.

It is also possible to construct bootstrap confidence intervals that are optimised for length, for a given level of coverage.

The coverage accuracy of bootstrap confidence intervals can be reduced by using the iterated bootstrap to estimate coverage error, and then adjust for it. Each application generally reduces coverage error by a factor of $n^{-1/2}$ in the one-sided case, and n^{-1} in the two-sided case. Usually, however, only one application is computationally feasible.

Other bootstrap confidence intervals (cont.)

Although the percentile-t approach has obvious advantages, these may not be realised in practice in the case of small samples. This is because bootstrapping the Studentised ratio involves simulating the ratio of two random variables, and unless sample size is sufficiently large to ensure reasonably low variability of the denominator in this expression, poor coverage accuracy can result.

Note too that percentile-t confidence intervals are not transformation-invariant, whereas intervals based on the percentile method are.

From some viewpoints, particularly that of good coverage performance in a very wide range of settings (an analogue of "robustness"), the most satisfactory approach is the coverage-corrected form (using the iterated bootstrap) of first type of percentile method interval, i.e. of $\hat{I}_{12}$ and $\hat{I}_{22}$ in one- and two-sided cases, respectively.

Bootstrap methods for time series

There are two basic approaches in the time-series case, applicable with or without a structural "model," respectively.

We shall say that we have a structural model for a time series, $X_1, \ldots, X_n$, if there is a continuous, deterministic method for generating the series from a sequence of independent and identically distributed "disturbances," $\epsilon_1, \epsilon_2,$ $\ldots$. The method should depend on a finite number of unknown, but estimable, parameters. Moreover, it should be possible to estimate all but a bounded number of the disturbances from n consecutive observations of the time series.

Bootstrap for time series with structural model

We call the model *structural* because the parameters describe only the structure of the way in which the disturbances drive the process. In particular, no assumptions are made about the disturbances, apart from standard moment conditions. In this sense the setting is nonparametric, rather than parametric.

The best known examples of structural models are those related to linear time series, for example the moving average

$$X_j = \mu + \sum_{i=1}^{p} \theta_i \, \epsilon_{j-i+1} \, ,$$

or an autoregression such as

$$X_j - \mu = \sum_{i=1}^{p} \omega_i \, (X_{j-i+1} - \mu) + \epsilon_j \, ,$$

where μ, $\theta_1, \ldots, \theta_p$, $\omega_1, \ldots, \omega_p$, and perhaps also p, are parameters that have to be estimated.

Bootstrap for time series with structural model (continued, 1)

In this setting the usual bootstrap approach to inference is as follows:

(1) Estimate the parameters of the structural model (e.g. μ and $\omega_1, \ldots, \omega_p$ in the autoregression example), and compute the residuals (i.e. "estimates" of the ϵ_j's), using standard methods for time series.

(2) Generate the "estimated" time series, in which true parameter values are replaced by their estimates and the disturbances are resampled from among the estimated ones, obtaining a bootstrapped time series $X_1^*, \ldots, X_n^*$, for example (in the autoregressive case)

$$X_j^* - \widehat{\mu} = \sum_{i=1}^{p} \widehat{\omega}_i \, (X_{j-i+1}^* - \widehat{\mu}) + \epsilon_j^* \, .$$

Bootstrap for time series with structural model (continued, 1)

(3) Conduct inference in the standard way, using the resample $X_1^*, \ldots, X_n^*$ thus obtained.

For example, to construct a percentile-t confidence interval for μ in the autoregressive example, let $\widehat{\sigma}^2$ be a conventional time-series estimator of the variance of $n^{1/2}\widehat{\mu}$, computed from the data $X_1, \ldots, X_n$; let $\widehat{\mu}^*$ and $(\widehat{\sigma}^*)^2$ denote the versions of $\widehat{\mu}$ and $\widehat{\sigma}^2$ computed from the resampled data $X_1^*, \ldots, X_n^*$; and construct the percentile-t interval based on using the bootstrap distribution of

$$T^* = n^{1/2}(\widehat{\mu}^* - \widehat{\mu})/\widehat{\sigma}^*$$

as an approximation to the distribution of

$$T = n^{1/2}(\widehat{\mu} - \mu)/\widehat{\sigma}\,.$$

Bootstrap for time series with structural model (continued, 2)

All the standard properties we have already noted, founded on Edgeworth expansions, apply without change provided the time series is sufficiently short-range dependent. Early work on theory in the structural time series case includes that of

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. Ann. Statist. 16, 1709–1722.

Bootstrap for time series with structural model (continued, 3)

It is common in this setting not to be able to "estimate" n disturbances ϵ_j, based on a time series of length n. For example, in the context of autoregressions we can generally estimate no more than $n - p$ of the disturbances. But this does not hinder application of the method; we merely resample from a set of $n - p$, rather than n, values of $\widehat{\epsilon}_j$.

Usually it is assumed that the disturbances have zero mean. We reflect this property empirically, by centring the $\widehat{\epsilon}_j$'s at their "sample" mean before resampling.

Bootstrap for time series without structural model

In some cases, for example where highly nonlinear filters have been applied during the process of recording data, it is not possible or not convenient to work with a structural model. There is a variety of bootstrap methods for conducting inference in this setting, based on "block" or "sampling window" methods. We shall discuss only the block bootstrap approach.

Block bootstrap for time series

Just as in the case of a structural time series, the block bootstrap aims to construct simulated versions "of" the time series, which can then be used for inference in a conventional way.

The method involves sampling blocks of consecutive values of the time series, say X_{I+1}, $\ldots, X_{I+b}$, where $0 \le I \le n-b$ is chosen in some random way; and placing them one after the other, in an attempt to reproduce the series. Here, b denotes block length.

Assume we can generated blocks $X_{I_j+1}, \ldots,$ X_{I_j+b}, for $j \ge 1$, *ad infinitum* in this way. Create a new time series, $X_1^*, X_2^*, \ldots$, identical to:

$$X_{I_1+1}, \ldots, X_{I_1+b}, X_{I_2+1}, \ldots, X_{I_2+b}, \ldots$$

The resample $X_1^*, \ldots, X_n^*$ is just the first n values in this sequence.

Block bootstrap for time series (continued, 1)

There is a range of methods for choosing the blocks. One, the “fixed block” approach, involves dividing the series $X_1, \ldots, X_n$ up into m blocks of b consecutive data (assuming $n = bm$), and choosing the resampled blocks at random. In this case the I_j’s are independent and uniformly distributed on the values $1, b+1, \ldots, (m-1)b+1$. The blocks in the fixed-block bootstrap do not overlap.

Another, the “moving blocks” technique, allows block overlap to occur. Here, the I_j’s are independent and uniformly distributed on the values $0, 1, \ldots, n-b$.

Block bootstrap for time series (continued, 2)

In this way the block bootstrap attempts to preserve exactly, within each block, the dependence structure of the original time series $X_1, \ldots, X_n$. However, dependence is corrupted at the places where blocks join.

Therefore, we expect optimal block length to increase with strength of dependence of the time series.

Techniques have been suggested for matching blocks more effectively at their ends, for example by using a Markovian model for the time series. This is sometimes referred to as the "matched block" bootstrap.

Difficulties with the block bootstrap

The main problem with the block bootstrap is that the block length, b, which is a form of smoothing parameter, needs to be chosen. Using too small a value of b will corrupt the dependence structure, increasing the bias of the bootstrap method; and choosing b too large will give a method which has relatively high variance, and consequent inaccuracy.

Another difficulty is that the percentile-t approach cannot be applied in the usual way with the block bootstrap, if it is to enjoy high levels of accuracy. This is because the corruption of dependence at places where adjacent blocks joins, significantly affects the relationship between the numerator and the denominator in the Studentised ratio, with the result that the block bootstrap does not effectively capture skewness. However, there are ways of removing this problem.

Successes of the block bootstrap

Nevertheless, the block bootstrap, and related methods, give good performance in a range of problems where no other techniques work effectively, for example inference for certain sorts of nonlinear time series.

The block bootstrap also has been shown to work effectively with spatial data. There, the blocks are sometimes referred to as "tiles," and either of the fixed-block or moving-block methods can be used.

References for block bootstrap

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann. Statist. 14, 1171–1179.

Hall, P. (1985). Resampling a coverage pattern. Stochastic Process. Appl. 20, 231–246.

Künsch, H.-R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Statist. 17, 1217–1241.

Politis, D.N., Romano, J.P., Wolf, M. (1999). Subsampling. Springer, New York.

Bootstrap in non-regular cases

There is a “meta theorem” which states that the standard bootstrap, which involves constructing a resample that is of (approximately) the same size as the original sample, works (in the sense of consistently estimating the limiting distribution of a statistic) if and only if that statistic’s distribution is asymptotically Normal.

It does not seem possible to formulate this as a general, rigorously provable result, but it nevertheless appears to be true.

The result underpins our discussion of bootstrap confidence regions, which has focused on the case where the statistic is asymptotically Normal. Therefore, rather than take up the issue of whether the bootstrap estimate of the statistic’s distribution is asymptotically Normal, we have addressed the problem of the size of coverage error.

Example of non-regular cases

Perhaps the simplest example where this approach fails is that of approximating the distributions of extreme values. To appreciate why there is difficulty, consider the problem of approximating the joint distribution of the two largest values of a sample, $X_1, \ldots, X_n$, from a continuous distribution. The probability that the two largest values in a resample, $X_1^*, \ldots, X_n^*$, drawn by sampling with replacement from the sample, both equal $\max X_i$, is

$$1 - \left(1 - \frac{1}{n}\right)^{n-1} - n\,\frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \to 1 - 2e^{-1}$$

as $n \to \infty$.

The fact that the probability does not converge to zero makes it clear that the joint distribution of the two largest values in the bootstrap sample cannot consistently estimate the joint distribution of the two largest data.

The m-out-of-n bootstrap

The most commonly used approach to overcoming this difficulty, in the extreme-value example and many other cases, is the *m-out-of-n bootstrap.* Here, rather than draw a sample of size n we draw a sample of size $m < n$, and compute the distribution approximation in that case. Provided

$$m = m(n) \to \infty \quad \text{and} \quad m/n \to 0$$

the m-out-of-n bootstrap gives consistent estimation in most, probably all, settings.

For example, this approach can be used to consistently approximate the distribution of the mean of a sample drawn from a very heavy-tailed distribution, for example one in the domain of attraction of a non-Normal stable law.

The m-out-of-n bootstrap (continued)

The main difficulty with the m-out-of-n bootstrap is choosing the value of m. Like block length in the case of the block bootstrap, m is a smoothing parameter; large m gives low variance but high bias, and small m has the opposite effect. In most problems where we would wish to apply the m-out-of-n bootstrap, it proves to be quite sensitive to selection of m.

A secondary difficulty is that the accuracy of m-out-of-n bootstrap approximations is not always good, even if m is chosen optimally. For example, when the m-out-of-n bootstrap is applied to distribution approximation problems, the error is often of order $m^{-1/2}$, which, since $m/n \to 0$, is an order of magnitude worse than $n^{-1/2}$.

Conclusion

Nevertheless, there is very substantial theoretical evidence that the bootstrap works quite well in a particularly wide range of statistical problems, and theoretical and empirical evidence that it performs very well indeed in some settings.

It is currently the only viable method for solving some problems, where asymptotic approximations are either not available, or are poor. (Certain extreme-value problems are of this type.)

For these reasons the bootstrap is a vital component of contemporary statistical methodology.