1 INTRODUCTION

1.1 SUMMARY

- **Bootstrap principle:** definition; history; examples of problems that can be solved; different versions of the bootstrap.
- **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
- **Bootstrap iteration:** principle and theory
- **Bootstrap in non-regular cases:** cases where the bootstrap is inconsistent; difficulties that the bootstrap has modelling extremes
- **Bootstrap for time series:** 'structural' and 'non-structural' implementations; block bootstrap methods
- **Bootstrap for nonparametric function estimation**

1.2 WHAT IS THE BOOTSTRAP?

The best known application of the bootstrap is to estimating the mean, μ say, of a population with distribution function F, from data drawn by sampling randomly from that population. Now,

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

The sample mean is the same functional of the empirical distribution function, i.e. of

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

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.

More generally, if $\theta_0 = \theta(F)$ denotes the true value of a parameter, where θ is a functional, then

$$\hat{\theta} = \theta(\widehat{F})$$

is the bootstrap estimator of θ_0.

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.

2 Prehistory of the bootstrap

2.1 Interpretation of 19th Century Contributions

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.

2.2 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.

Arguably the first person to be involved in this type of work was not a statistician but an Indian Civil Servant, John Hubback. Hubback, an 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.

In that work he introduced a version of the block bootstrap for spatial data, in the form of crop yields in fields scattered across parts of Bihar and Orissa.

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.

Hubback's research 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)

2.3 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, and the earlier work of Hubback, anticipated the much more modern technique of the block bootstrap.

2.4 Contributions in 1950s and 1960s

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).

2.5 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.

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 their economic well-being and their self-esteem.

3 Efron's bootstrap

3.1 Overview of 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 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 combined 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.

3.2 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 $F_1 = \widehat{F}$.

Example 1: bias correction

Here, $\theta = \theta(F_0)$ is the true value of a parameter, and $\hat{\theta} = \theta(F_1)$ is its estimator; t is an additive adjustment to $\hat{\theta}$; $\hat{\theta} + t$ 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 $\hat{\theta}$, 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 $E(\hat{\theta} - \theta + t) = 0$, which is equivalent to (1).

Example 2: confidence interval

Here we take

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

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

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

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.

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

3.4 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.

3.5 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

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

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

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

where the subscript bc denotes 'bias corrected.'

Sometimes we can compute $E(\hat{\theta}^* \mid \widehat{F})$ directly, but in many instances we can access it only through numerical approximation. For example, conditional on $\mathcal{X}$, we can compute independent values $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ of $\hat{\theta}^*$, and take

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*_b$$

to be our numerical approximation to $E(\hat{\theta}^* \mid \widehat{F})$.

3.6 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,

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

Since $\hat{\theta}$, 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.

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

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

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

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

For example, it might denote the largest t such that

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

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,

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

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 $B_0 \leq B \leq \infty$, 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.

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 $(-\infty, \hat{\theta} + \hat{t}\,]$, where $t = \hat{t}$ is the (approximate) solution of the equation

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

(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 $\hat{\theta}$ 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 $\hat{\theta}^*_{(\nu)}$ denote the νth largest of the B simulated values of $\hat{\theta}^*$, and let the confidence interval be $(-\infty, \hat{\theta}^*_{(\nu)}]$. Then,

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

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

3.7 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.

- 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}

EXERCISE: Derive the formula $N(n) \sim (n\pi)^{-1/2}\, 2^{2n-1}$, and the table above.

- 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}

EXERCISE: Show that $\mathcal{X}$ is the most likely resample to be drawn, and derive the formulae $p_n = n!/n^n \sim (2n\pi)^{1/2}\, e^{-n}$ and the table above.

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

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

where

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

is the bootstrap form of the empirical distribution function. (The bootstrap resample 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 $T(F_1)$:

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

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

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

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 $\hat{t}$ for $T(F_1)$,

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

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

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

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

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

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

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

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

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

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

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

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}:

$$P\{\theta(F_1) - \hat{t} \leq \theta(F_0) \leq \theta(F_1) + \hat{t}\} \;=\; 1 - \alpha + O(n^{-1})\,.$$

That is,

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

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 $O(n^{-1})$.'

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

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

then

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

That is, 'the coverage error of the nominal $1-\alpha$ level, one-sided percentile bootstrap confidence interval $(-\infty, \hat{\theta}+\hat{t}]$ equals $O(n^{-1/2})$.'

4.1 WHY BOTHER WITH THE BOOTSTRAP?

It can be shown that the orders of magnitude of error discussed above are identical to those associated with intervals based on conventional normal approximations. That is, standard asymptotic-theory confidence intervals, constructed by appealing to the central limit theorem, cover the unknown parameter with a given probability plus an error that equals $O(n^{-1})$ in the case of two-sided intervals, and $O(n^{-1/2})$ for their one-sided counterparts. What has been gained?

In fact, there are several advantages in using the bootstrap. First, the percentile bootstrap confidence interval for θ does not require a variance estimator. However, its 'asymptotic theory' counterpart requires us to compute an estimator, $\hat{\sigma}^2$, of the asymptotic variance σ^2 of $n^{1/2}\,(\hat{\theta}-\theta)$, and in non-standard problems this can be a considerable challenge. In effect, the percentile-bootstrap computes the variance estimator for us, implicitly, without our having to work out the value.

Secondly, there are several ways of improving a percentile-bootstrap interval so as to reduce the order of magnitude of coverage error without making its calculation significantly more difficult. One of these is the method of bootstrap iteration, which we shall consider next. In a variety of respects, iteration of a standard percentile-bootstrap con-

fidence interval is the most appropriate approach to constructing confidence intervals. For example, it reduces the level of coverage error by an order of magnitude, relative to either the standard percentile method or its asymptotic-theory competitor, and it does not require variance calculation.

5 Bootstrap Iteration

5.1 Basic principle behind 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.

5.2 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 be solved only by Monte Carlo simulation: calculating $\hat{t}$ 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.

5.3 IMPLEMENTING THE DOUBLE BOOTSTRAP

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

$$P(\theta \leq \hat{\theta} + t) = 1 - \alpha\,,$$

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

One application of the bootstrap involves creating resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$; computing the

version, $\hat{\theta}_b^*$, of $\hat{\theta}$ from $\mathcal{X}_b^*$; and choosing $t = \hat{t}$ such that

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

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

For the next application of the bootstrap, from each resample $\mathcal{X}_b^*$ we draw C re-resamples, $\mathcal{X}_{b1}^{**}, \ldots, \mathcal{X}_{bC}^{**}$, the cth (for $1 \leq c \leq C$) 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, $\hat{\theta}_{bc}^{**}$, of $\hat{\theta}$ from $\mathcal{X}_{bc}^{**}$, and choose $t = \hat{t}_b^*$ such that

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

as nearly as possible.

Interpret $\hat{t}_b^*$ as the version of $\hat{t}$ 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

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

as nearly as possible.

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

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

5.4 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, Student's t bootstrap

one-sided confidence intervals that have coverage error $O(n^{-1})$. Application of the double bootstrap to them reduces the order of their coverage error to order $n^{-1/2} \times n^{-1} = n^{-3/2}$.

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 order n^{-2}.

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

5.5 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.

5.6 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, $\hat{t} = T(F_1)$, is given by

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

and that the bias-corrected form of the estimator $\theta(F_1)$ is

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

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

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

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 $F_1, \ldots, F_{j+1}$.

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

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 $\hat{\theta}$ (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.

5.7 ITERATED BOOTSTRAP FOR BIAS VARIANCE ESTIMATION

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

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

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

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

However, if we correct for bias multiplicatively rather than additively, a single application of bias correction produces the unbiased estimator S^2.

EXERCISE: Derive these results.

6 PERCENTILE-*t* CONFIDENCE INTERVALS

6.1 DEFINITION AND BASIC PROPERTIES

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 estimators 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.

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

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

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,

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

where the approximations derive from the central limit theorem. (We could take t_α to be the α-level quantile of the standard normal distribution, in which case the second approximation is an identity.) Hence, $(-\infty, \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,t_\alpha]$ is an approximate $(1-\alpha)$-level confidence interval for θ.

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

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

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

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

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

Recall that the Normal-approximation confidence interval for θ was

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

where t_α is the α-level quantile of the standard normal distribution. If we replace the confidence interval endpoint here by its percentile bootstrap version, considered earlier, we obtain a percentile bootstrap confidence for which the coverage error is generally of size $n^{-1/2}$.

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

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

6.2 Comparison with normal approximation

Standard one-sided confidence intervals based on the normal approximation have coverage error of order $n^{-1/2}$. This is the level ensured by the Berry-Esseen theorem, and generally cannot be improved unless the sampling distribution has symmetry properties. (However, two-sided confidence intervals based on the percentile method have coverage error of order n^{-1}, rather than $n^{-1/2}$.)

Note that, in contrast, the one-sided percentile-t interval has coverage error of order n^{-1}. (It's two-sided version has the same order of coverage, not $O(n^{-1})$.)

7 MOMENTS AND CUMULANTS

7.1 DEFINITIONS

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 $\log \chi(t)$:

$$\chi(t) \;=\; \exp\left\{\kappa_1\, it + \tfrac{1}{2}\,\kappa_2\,(it)^2 + \ldots + \tfrac{1}{j!}\,\kappa_j\,(it)^j + \ldots\right\}\,.$$

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

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

7.2 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.

7.3 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

$$\chi_n(t) = E \left\{ \exp(itS_n) \right\}$$

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

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

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

Now expand the exponent, collecting together terms of order $n^{-j/2}$ for each $j \geq 0$:

$$
\chi_n(t) = e^{-t^2/2}\,\left\{1 + n^{-1/2}\,r_1(it) + \ldots + n^{-j/2}\,r_j(it) + \ldots\right\} \,,
$$

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,

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

Note that none of the polynomials has a constant term.

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

7.4 EXPANSION OF DISTRIBUTION FUNCTION

Rewrite the expansion as:

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

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,

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

where

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

7.5 FORMULA FOR R_j, PART I

Integration by parts gives:

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

$$= (-it)^{-j} \int_{-\infty}^{\infty} e^{itx}\, d\Phi^{(j)}(x)\,,$$

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,

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

7.6 HERMITE POLYNOMIALS

The Hermite polynomials,

$$\begin{aligned}
&\mathrm{He}_0(x) = 1\,, \quad \mathrm{He}_1(x) = x\,,\\
&\mathrm{He}_2(x) = x^2 - 1\,,\\
&\mathrm{He}_3(x) = x\,(x^2 - 3)\,,
\end{aligned}$$

$$\mathrm{He}_4(x) = x^4 - 6\,x^2 + 3\,,$$
$$\mathrm{He}_5(x) = x\,(x^4 - 10\,x^2 + 15)\,, \ldots$$

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,

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

7.7 Formula for R_j, part II

Therefore, if

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

(recall that none of the polynomials r_j has a constant term), then

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

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 $j+2$.

Examples:

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

EXERCISE: Derive these formulae. [Hint: This is straightforward, given what we have proved already.]

7.8 ASYMPTOTIC EXPANSIONS

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

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

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 holds rarely for distributions which are not very closely connected to the Normal distribution.

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:

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

A sufficient regularity condition for this result is

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

A rigorous derivation of the expansion under these restrictions was given first by Harald Cramér.

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

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. if, 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.

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

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

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

The 'gap' between cases where Cramér's continuity condition holds, and the case where X has a lattice distribution, is well understood only for $j = 1$. It was shown by Esseen (1945) that the expansion

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

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

8 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 $f_n(x) = (d/dx)\, P(S_n \leq x)$, we have:

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

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

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

$$P(S_n = x) \;=\; n^{-1/2} \left\{ \phi(x) + n^{-1/2}\, R_1'(x) + \ldots + n^{-j/2}\, R_j'(x) + o(n^{-j/2}) \right\},$$

uniformly in points x in the support of the distribution.

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

EXERCISE: Derive the version of this local lattice expansion in the case where the unstandardised form of X has the Binomial Bi(m, 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 $\binom{n}{r}$.]

9 Expansions in More General Cases

9.1 Contribution by Bhattacharya and Ghosh (1978)

The year 2008 was the 80th 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.

Bhattacharya and Ghosh dealt with quantities 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 $t = (t^{(1)}, \ldots, t^{(d)})^{\mathrm{T}}$ denote a d-vector, and let $\chi(t) = E\{\exp(it^{\mathrm{T}}X)\}$ 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.

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

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

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

Then,

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

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.

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:

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

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:

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

The functions R_j appearing here are exactly those obtained by, first, finding the cumulant expansion of the distribution of $Q_n(\bar{X} - \mu)$, then writing down formally its inverse as a formula for the distribution of $Q_n(\bar{X} - \mu)$.

Note that at this point we are only developing a conjectured formula for the distribution of $Q_n(\bar{X} - \mu)$, not deriving its validity. The validity of the expansion is given by Bhattacharya and Ghosh's theorem.

9.3 Polynomials in Studentised and non-Studentised cases

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

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

We know already that in the non-Studentised case,

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

10 Cornish-Fisher Expansions

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

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

This is an expansion of a probability for a given value of 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:

$$\xi_\alpha \;=\; z_\alpha + n^{-1/2}\,P_1^{\rm cf}(z_\alpha) + \ldots + \; n^{-j/2}\,P_j^{\rm cf}(z_\alpha) + o(n^{-j/2})\,,$$

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

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

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

etc.

EXERCISE: Derive these formulae.

11 STUDENTISED AND NON-STUDENTISED ESTIMATORS

11.1 INTRODUCTION

Let $\hat{\theta} = \theta(\widehat{F})$ denote the bootstrap estimator of a parameter $\theta = \theta(F)$, 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

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

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

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

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

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.

11.2 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:

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

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

11.3 BOOTSTRAP EDGEWORTH EXPANSIONS

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

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

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:

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

Under moment and smoothness conditions, as in the theorem of Bhattacharya and Ghosh stated earlier, these expansions are valid uniformly, and in probability and with probability 1, up to remainders that are of strictly smaller order than the last included term:

$$\sup_{-\infty<x<\infty} \left| P(S^* \le \hat\sigma x \mid \mathcal{X}) - \left\{ \Phi(x) + n^{-1/2}\, \widehat{P}_1(x)\, \phi(x) + \ldots + n^{-j/2}\, \widehat{P}_j(x)\, \phi(x) \right\} \right| = o_p\big(n^{-j/2}\big)\,,$$

$$\sup_{-\infty<x<\infty} \left| P(T^* \le x \mid \mathcal{X}) - \left\{ \Phi(x) + n^{-1/2}\, \widehat{Q}_1(x)\, \phi(x) + \ldots + n^{-j/2}\, \widehat{Q}_j(x)\, \phi(x) \right\} \right| = o_p\big(n^{-j/2}\big)\,,$$

with probability 1.

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.

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

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

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

$$\begin{aligned}
\hat\kappa_3 &= \hat\sigma^{-3}\,n^{-1}\sum_{i=1}^n (X_i-\bar X)^3\,,\\
\hat\kappa_4 &= \hat\sigma^{-4}\,n^{-1}\sum_{i=1}^n (X_i-\bar X)^4 - 3\,,
\end{aligned}$$

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

Recall that we term 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 were 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.)

Of course, in many practical settings the terms "skewness" and "kurtosis" are used more loosely than in the case of a sum of independent random variables, where skewness is defined to be the third central moment and kurtosis is the fourth central moment minus three times the square of the variance. However, the principle is the same — the term of size $n^{-1/2}$ represents a correction for asymmetry, and the term of size n^{-1} is a correction for tail weight and a second-order correction for asymmetry.

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

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

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.

14 ACCURACY OF BOOTSTRAP APPROXIMATIONS

It follows that

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

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}:

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

$$P(T^* \le x \mid \mathcal{X}) \;=\; P(T \le x) + O_p(n^{-1})\,.$$

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

These orders of approximation are valid uniformly in x.

These results underpin the performance of bootstrap methods in distribution approximation, and show their 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 use effectively approximations based on S^* we generally 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^*.

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

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

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

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

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

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

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

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

$$\widehat{P}_2^{\rm cf}(x) = \widehat{P}_1(x)\, \widehat{P}_1'(x) - \tfrac{1}{2}\, x\, \widehat{P}_1(x)^2 - \widehat{P}_2(x) \,,$$

$$\widehat{Q}_2^{\rm cf}(x) \;=\; \widehat{Q}_1(x)\,\widehat{Q}_1'(x) - \tfrac{1}{2}\,x\,\widehat{Q}_1(x)^2 - \widehat{Q}_2(x)\,,$$

etc.

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

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

and hence that

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

(These orders of approximation are valid uniformly in $\alpha \in [\epsilon, 1-\epsilon]$ for any $\epsilon \in (0, \frac{1}{2})$.)

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.

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

The intervals

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

cover θ with probability exactly α, but are in general not computable, since we do not know either ξ_α or η_α. On the other hand, their bootstrap counterparts

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

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.

The "other" percentile confidence intervals for θ are

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

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 $\hat{I}_{12}$, $\hat{J}_1$, K_1 and $\widehat{K}_1$ to converge to $1-\alpha$ as $n \to \infty$. However, the convergence rate is generally only $n^{-1/2}$. The exceptions are $\hat{I}_{11}$ and $\hat{J}_1$, for which coverage error equals $1 - \alpha + O(n^{-1})$.

The interval $\hat{I}_{11}$ 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 $\hat{J}_1$ is both useful and has good coverage accuracy.

17 Advantages of Using a Pivotal Statistic

17.1 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.

17.2 Derivation of these properties

We begin with the case of the confidence interval

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

our aim being to show that

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

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

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

and that, by Cornish-Fisher expansion,

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

where η_α is defined by $P(T \leq \eta_\alpha) = \alpha$.

Therefore,

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

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

$$P(T \leq x) \;=\; \Phi(x) + n^{-1/2}\,Q_1(x)\,\phi(x) + O(n^{-1})\,,$$

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

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

$$= \; P(T \leq \eta_\alpha) + O(n^{-1}) \, .$$

(These steps need only Taylor expansion.) Therefore,

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

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

Hence,

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

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

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

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

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

$$N_1 = (-\infty, \hat{\theta} - n^{-1/2}\,\hat{\sigma}\,z_\alpha) = (-\infty, \hat{\theta} + n^{-1/2}\,\hat{\sigma}\,z_{1-\alpha})\,,$$

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

EXERCISE: Prove that

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

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 $\hat{J}_1$.

Similarly it can be proved that coverage error of the interval

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

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

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

However, this result fails for the more practical interval

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

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

Recall that $\hat{\xi}_\alpha$ is defined by

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

and that, by a Cornish-Fisher expansion,

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

where ξ_α is defined by $P(S \le \sigma\,\xi_\alpha) = \alpha$. Therefore,

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

Assuming, as before, that the remainder $O_p(n^{-1})$ can be treated as deterministic, we have, by Taylor expansion,

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

$$
\begin{aligned}
&= P(T \le \xi_\alpha) + O(n^{-1}) \\
&= P(T \le \eta_\alpha) + (\xi_\alpha - \eta_\alpha)\,\phi(\eta_\alpha) + O(n^{-1})\,,
\end{aligned}
$$

where we have used the fact that $\xi_\alpha - \eta_\alpha = O(n^{-1/2})$. Indeed,

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

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

Hence,

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

Therefore,

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

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.

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

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

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

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

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

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

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, as we now show.

Recall that one form of percentile-method confidence interval for θ is

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

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

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

20.1 COVERAGE OF TWO-SIDED PERCENTILE INTERVALS

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

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

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

$$P_1(z_{1-(\alpha/2)}) - Q_1(z_{1-(\alpha/2)}) = P_1(z_{\alpha/2}) - Q_1(z_{\alpha/2})\,.$$

Therefore,

$$\begin{aligned}P\{\theta \in \hat{I}_{22}(\alpha)\} &= P\left[\theta \in \hat{I}_{12}\{1-(\alpha/2)\}\right] - P\left[\theta \in \hat{I}_{12}(\alpha/2)\right] \\ &= 1 - (\alpha/2) + n^{-1/2}\{P_1(z_{1-(\alpha/2)}) - Q_1(z_{1-(\alpha/2)})\}\,\phi(z_{1-(\alpha/2)})\end{aligned}$$

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

Hence, 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

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

Its one- and two-sided forms have coverages given by:

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

EXERCISE: (1) Derive the latter two formulae. (2) Show that, when computing two-sided 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.)

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 $\hat{\theta}$, 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 $\hat{\theta}$ 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.

21 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.

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 range-respecting or 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 the percentile method interval, i.e. of $\hat{I}_{12}$ and $\hat{I}_{22}$ in one- and two-sided cases, respectively.

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 smooth, deterministic model for generating the series from a sequence of independent and identically distributed "disturbances," $\epsilon_1, \epsilon_2, \ldots$. The model 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.

22.1 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 a moving average

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

or an autoregression such as

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

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

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)

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

(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 $\hat{\sigma}^2$ be a conventional time-series estimator of the variance of $n^{1/2}\hat{\mu}$, computed

from the data $X_1, \ldots, X_n$; let $\hat{\mu}^*$ and $(\hat{\sigma}^*)^2$ denote the versions of $\hat{\mu}$ and $\hat{\sigma}^2$ computed from the resampled data $X_1^*, \ldots, X_n^*$; and construct the percentile-t interval based on using the bootstrap distribution of

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

as an approximation to the distribution of

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

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.

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 $\hat{\epsilon}_j$.

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

22.2 Bootstrap for time series without structural model: The block bootstrap

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.

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 $0 \leq I \leq n - b$ 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 generate blocks $X_{I_j+1}, \ldots, X_{I_j+b}$, for $j \geq 1$, *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.

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 $1, \ldots, n-b$.

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.

22.3 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 join, 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 overcoming this problem.

22.4 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.

22.5 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.

23 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 largely correct.

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.

23.1 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

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

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 a bootstrap sample cannot consistently estimate the joint distribution of the two largest data.

23.2 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 that

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

the m-out-of-n bootstrap gives consistent estimation in most, and possibly 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 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 $n^{-1/2}$.

24 Bootstrap Methods in Linear Regression

24.1 Regression model

Assume we observe pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, generated by the model

$$Y_i = g(x_i) + \epsilon_i \,, \tag{1}$$

where g is a function that might be determined either parametrically or nonparametrically, and the errors ϵ_i have zero mean.

In the study of regression we take the explanatory variables x_i to be fixed, either because they are pre-determined (e.g. were regularly spaced) or are conditioned upon.

In this case, the only source of randomness in the model is the errors, ϵ_i, and so it is those that we resample, in form of residuals, when implementing the bootstrap. Our choice of lower-case notation for the explanatory variables reflects this view.

24.2 Correlation model

Alternatively, we might take the view that the explanatory variables are genuinely random and must be treated us such. For example, the data pairs (X_i, Y_i), for $1 \leq i \leq n$, might be drawn by sampling randomly from a bivariate distribution, and in our analysis of those data we might wish to preserve all the implications of this randomness, rather than condition some of it away by regarding the X_i's as fixed.

This approach to analysis might be termed the study of correlation, rather than regression. It would be addressed in bootstrap terms by resampling the pairs (X_i, Y_i), rather than resampling the residuals.

It is important to appreciate that these two different approaches to resampling — sampling the residuals or sampling the pairs (X_i, Y_i), respectively — are appropriate in different settings, for different models. They are not alternative ways of doing the same thing, and can lead to different conclusions.

For example, results derived under the correlation model generally reflect the higher degree of variability present there; a part of this variability is "conditioned out" under the regression model. In particular, confidence and prediction intervals are generally a little wider under the correlation model than under the regression model.

24.3 PARAMETRIC REGRESSION

The good properties of percentile-t methods carry over to regression problems. However, in the setting of slope estimation those properties are significantly enhanced, and even the standard percentile method can perform unusually well.

For example, one-sided percentile-t confidence regions for slope have coverage error $O(n^{-3/2})$, not $O(n^{-1})$; and the error is only $O(n^{-2})$ in the case of two-sided intervals.

One-sided, standard percentile-method confidence intervals for slope, based on approximating the distribution of $\hat{\theta} - \theta$ by the conditional distribution of $\hat{\theta}^* - \hat{\theta}$, have coverage error $O(n^{-1})$ rather than the usual $O(n^{-1/2})$.

24.4 General definition of slope

Although these exceptional coverage properties apply only to estimates of slope, not to estimates of intercept parameters or means, slope may be interpreted very generally.

For example, in the polynomial regression model

$$Y_i = c + x_i\, d_1 + \ldots + x_i^m\, d_m + \epsilon_i\,,$$

where we observe (x_i, Y_i) for $1 \le i \le n$, we regard each d_j as a slope parameter. A one-sided percentile-t interval for d_j has coverage error $O(n^{-3/2})$, although a one-sided percentile-t interval for c or for

$$E(Y \,|\, x = x_0) = c + x_0\, d_1 + \ldots + x_0^m\, d_m$$

has coverage error of size n^{-1}.

24.5 Why is slope favoured especially?

The reason for good performance in the case of slope parameters is the extra symmetry conferred by design points. Note that, in the polynomial regression case, we may write

the model equivalently as

$$Y_i = c' + (x_i - \xi_1)\, d_1 + \ldots + \left(x_i^m - \xi_m\right) d_m + \epsilon_i \,,$$

where $\xi_j = n^{-1} \sum_i x_i^j$ and $c' = c + \xi_1\, d_1 + \ldots + \xi_m\, d_m$.

The extra symmetry arises from the fact that

$$\sum_{i=1}^{n} \left(x_i^j - \xi_j\right) = 0$$

for $1 \leq j \leq m$.

24.6 CORRELATION MODEL

The results implying good coverage accuracy hold under the regression model, but not necessarily for the correlation model. For example, in the case of the linear correlation model, the symmetry discussed above will persist provided that

$$E\left\{\sum_{i=1}^{n} (X_i - \xi_1)\, \epsilon_i^k\right\} = 0 \tag{1}$$

for sufficiently large k. Now, $\epsilon_i = Y_i - g(X_i)$, and as a result, (1) generally will not hold for $k \geq 1$. This means that, under the correlation model, the conventional properties of bootstrap confidence intervals hold; the special properties noted for regression, when estimating slope parameters, are not valid.

However, if the errors ϵ_i are independent of the explanatory variables X_i then (1) will hold for each k, and in such cases the enhanced features of the regression problem persist under the correlation model.

24.7 SIMPLE LINEAR REGRESSION

Consider the regression model,

$$Y_i = c' + x_i\, d + \epsilon_i = c + (x_i - \bar{x})\, d + \epsilon_i\,,$$

where the ϵ_i's are independent and identically distributed with zero mean and finite variance σ^2, $c' = c + d\,\bar{x}$ denotes the intercept, and d is the slope. (It turns out to be notationally simpler, at this point, to interchange the notations c and c'.) Define $\sigma_x^2 = n^{-1} \sum_i (x_i - \bar{x})^2$,

$$\hat{d} = \sigma_x^{-2}\, \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})\, Y_i$$

and $\hat{c} = \bar{Y} - \bar{x}\, \hat{d}$. Estimate $y_0 = E(Y \mid x = x_0) = c + x_0\, d$ by

$$\hat{y}_0 = \hat{c} + x_0\, \hat{d}\,,$$

and put

$$\begin{aligned} \sigma_y^2 &= 1 + \sigma_x^{-2}\, (x_0 - \bar{x})^2\,, \\ \hat{\epsilon}_i &= Y_i - \bar{Y} - (x_i - \bar{x})\, \hat{d} \end{aligned}$$

and $\hat{\sigma}^2 = n^{-1} \sum_i \hat{\epsilon}_i^2$, the latter estimating σ^2.

The asymptotic variances of $\hat{d}$ and $\hat{y}_0$ equal $\sigma^2/(n\,\sigma_x^2)$ and $\sigma^2\,\sigma_y^2/n$, respectively, and so $n^{1/2}\,(\hat{d}-d)\,\sigma_x/\hat{\sigma}$ and $n^{1/2}(\hat{y}_0-y_0)/(\hat{\sigma}\sigma_y)$ are asymptotically pivotal.

24.8 BOOTSTRAPPING THE SIMPLE LINEAR REGRESSION MODEL

The residuals,

$$\hat{\epsilon}_i = Y_i - \bar{Y} - (x_i - \bar{x})\,\hat{d}\,,$$

are centred, in that $\sum_i \hat{\epsilon}_i = 0$. Therefore we may resample randomly, with replacement, from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, obtaining $\epsilon_1^*, \ldots, \epsilon_n^*$ say; and take as our bootstrap resample the pairs $(x_1, Y_1^*), \ldots, (x_n, Y_n^*)$, where

$$Y_i^* = \hat{c} + (x_i - \bar{x})\,\hat{d} + \epsilon_i^*\,.$$

Note particularly that, reflecting the fact that we condition upon the explanatory variables, those quantities are the same as in the original dataset.

In regression problems where the residuals are not centred, for example in nonparametric regression, we generally centre them, for example by subtracting their average value, before resampling.

24.9 ESTIMATING QUANTILES OF DISTRIBUTION OF $\hat{d}$

Let $\hat{c}^*$, $\hat{d}^*$ and $\hat{\sigma}^*$ have the same formulae as $\hat{c}$, $\hat{d}$ and $\hat{\sigma}$, respectively, except that we replace Y_i by Y_i^* throughout. The bootstrap versions of

$$S \;=\; n^{1/2}\,(\hat{d}-d)\,\sigma_x/\sigma\,,$$

$$T = n^{1/2}(\hat{d} - d)\,\sigma_x/\hat{\sigma}$$

are

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

respectively.

We estimate the quantiles ξ_α and η_α of the distributions of S and T by $\hat{\xi}_\alpha$ and $\hat{\eta}_\alpha$, respectively, where

$$P(S^* \le \hat{\xi}_\alpha \,|\, \mathcal{X}) = \alpha\,, \quad P(T^* \le \hat{\eta}_\alpha \,|\, \mathcal{X}) = \alpha\,,$$

and $\mathcal{X} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$ denotes the dataset.

24.10 BOOTSTRAP CONFIDENCE INTERVALS FOR d

One-sided bootstrap confidence intervals for d, with nominal coverage α, are given by

$$\begin{aligned} \hat{I}_{11} &= \left(-\infty, \hat{d} - n^{-1/2}(\sigma/\sigma_x)\,\hat{\xi}_{1-\alpha}\right)\,, \\ \hat{I}_{12} &= \left(-\infty, \hat{d} - n^{-1/2}(\hat{\sigma}/\sigma_x)\,\hat{\xi}_{1-\alpha}\right)\,, \\ \hat{J}_1 &= \left(-\infty, \hat{d} - n^{-1/2}(\hat{\sigma}/\sigma_x)\,\hat{\eta}_{1-\alpha}\right)\,. \end{aligned}$$

These are direct analogues of the intervals $\hat{I}_{11}$, $\hat{I}_{12}$ and $\hat{J}_1$ introduced earlier in non-regression problems. In particular, $\hat{I}_{12}$ and $\hat{J}_1$ are standard percentile-method and percentile-

t bootstrap confidence regions.

Following the line of argument given earlier, we would expect $\hat{I}_{11}$ and $\hat{J}_1$ to have coverage error $O(n^{-1})$. In fact, they both have coverage error equal to $O(n^{-3/2})$. However, $\hat{I}_{11}$ is not of practical use, since it depends on the unknown σ, so we shall not treat it any further.

Likewise, we would expect $\hat{I}_{12}$ to have coverage error of size $n^{-1/2}$. However, we shall show that the error is actually of order n^{-1}.

Note that, although $\hat{I}_{12}$ involves the variance estimator $\hat{\sigma}$, it can be constructed numerically without resorting to computing $\hat{\sigma}$.

Indeed, $\hat{I}_{12}$ is identical to the interval

$$\hat{I}_{12} = \left(-\infty, \hat{d} - \widehat{w}_{1-\alpha}\right) ,$$

where $\widehat{w}_{1-\alpha}$ is the standard percentile-method estimator of $w_{1-\alpha}$, the latter defined by

$$P(\hat{d} - d \leq w_{1-\alpha}) = 1 - \alpha \, .$$

In particular, $\widehat{w}_{1-\alpha}$ is defined by

$$P(\hat{d}^* - \hat{d} \leq \widehat{w}_{1-\alpha} \,|\, \mathcal{X}) = 1 - \alpha \, .$$

The interval $\hat{J}_1$ is a standard percentile-t bootstrap confidence interval.

24.11 Polynomials in Edgeworth expansions

Edgeworth expansions for the non-Studentised and Studentised statistics, S and T respectively, are given by:

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

where

$$P_1(u) = Q_1(u) = \tfrac{1}{6}\,\gamma\,\gamma_x\,(1-u^2)\,,$$

$\gamma = E(\epsilon/\sigma)^3$, $\gamma_x = n^{-1}\sum_i \{(x_i - \bar{x})/\sigma_x\}^3$, and P_2 and Q_2 are odd, quintic polynomials, with

$$P_2(u) = Q_2(u) + u\left\{2 + (3/24)\,(2-\kappa)\left(u^2-3\right)\right\}$$

and $\kappa = E(\epsilon/\sigma)^4 - 3$.

Note particularly that $P_1 = Q_1$.

24.12 Why does P_1 equal Q_1?

To understand why, it is helpful to treat S as an approximation to T. Indeed, note that, by definition of $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\epsilon_i - \bar{\epsilon} - (x_i - \bar{x})\,(\hat{d} - d)\right\}^2 \\
&= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2) + O_p(n^{-1})\,.
\end{aligned}
$$

Therefore, defining

$$
\Delta = \tfrac{1}{2}\,n^{-1}\,\sigma^{-2}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2)\,,
$$

and recalling that

$$
\begin{aligned}
S &= n^{1/2}\,(\hat{d} - d)\,\sigma_x/\sigma\,, \\
T &= n^{1/2}\,(\hat{d} - d)\,\sigma_x/\hat{\sigma}\,,
\end{aligned}
$$

we deduce that

$$
T = S\,(1 - \Delta) + O_p(n^{-1})\,. \tag{1}
$$

Making use of this approximation, the symmetry property

$$
\sum_{i=1}^{n}(x_i - \bar{x}) = 0\,, \tag{2}
$$

and the representation

$$
S = n^{-1/2}\,\sigma_x^{-1}\,\sigma^{-1}\sum_{i=1}^{n}(x_i - \bar{x})\,\epsilon_i\,,
$$

it is readily proved that

$$E\{S\,(1-\Delta)\}^j = E(S^j) + O(n^{-1}) \tag{3}$$

for $j = 1, 2, 3$.

EXERCISE: Derive (3) for $j = 1, 2, 3$.

Therefore, the first three cumulants of S and $S\,(1-\Delta)$ agree up to and including terms of order $n^{-1/2}$.

It follows that Edgeworth expansions of the distributions of S and $S\,(1-\Delta)$ differ only in terms of order n^{-1}. In view of (1), the same is true of the distributions of S and T:

$$P(S \le u) = P(T \le u) + O(n^{-1})\,.$$

Therefore, the $n^{-1/2}$ terms in the expansions must be identical; that is, $P_1 = Q_1$.

The chief ingredient in this argument is the symmetry property (2). It, and its analogues, guarantee that, in the problem of slope estimation for general regression problems, $P_1 = Q_1$.

24.13 CONSEQUENCES OF THE PROPERTY $P_1 = Q_1$

The identity $P_1 = Q_1$ implies that, to first order (i.e. up to and including terms of order $n^{-1/2}$), estimating the distribution of S is the same as estimating the distribution of T.

As we saw earlier in non-regression problems, the percentile method estimates the distribution of S, whereas the percentile-t method estimates the distribution of T. The fact that, in the setting of estimating slope in regression, $P_1 = Q_1$, means that these two techniques give the same results up to and including terms of order $n^{-1/2}$. They differ only in terms of order n^{-1}, and terms of higher order.

Therefore, since one-sided confidence intervals based on the percentile-t method have coverage error equal to $O(n^{-1})$, the same must be true for confidence intervals based on the percentile method.

24.14 PROPERTIES OF PERCENTILE-t CONFIDENCE REGIONS

The coverage error of a one-sided percentile-t confidence interval for d is of order $n^{-3/2}$, rather than the usual n^{-1}. That is, with $\hat{J}_1$ defined by

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

it can be shown that

$$P\big(d \in \hat{J}_1\big) = 1 - \alpha + O(n^{-3/2}) \, . \tag{4}$$

Since terms of odd order in $n^{-3/2}$ cancel from formulae for coverage error of two-sided confidence intervals, then the two-sided, percentile-t bootstrap confidence interval for d has coverage error of order n^{-2}, rather than the usual n^{-1}.

24.15 Derivation of (4)

A proof of (4) can be given as follows. Recall that

$$P(T \le u) = \Phi(u) + n^{-1/2}\, Q_1(u)\, \phi(u) + \, n^{-1}\, Q_2(u)\, \phi(u) + \dots \,,$$

where

$$Q_1(u) = \tfrac{1}{6}\, \gamma\, \gamma_x\, (1 - u^2)\,, \quad \gamma = E(\epsilon/\sigma)^3\,, \quad \gamma_x = n^{-1} \sum_{i=1}^{n} \{(x_i - \bar{x})/\sigma_x\}^3$$

The bootstrap version of the expansion is

$$P(T^* \le u \,|\, \mathcal{X}) = \Phi(u) + n^{-1/2}\, \widehat{Q}_1(u)\, \phi(u) + \, n^{-1}\, \widehat{Q}_2(u)\, \phi(u) + \dots \,,$$

where

$$\widehat{Q}_1(u) = \tfrac{1}{6}\, \hat{\gamma}\, \gamma_x\, (1 - u^2)$$

and $\hat{\gamma} = E\{(\hat{\epsilon}_i^*)^3 \,|\, \mathcal{X}\}/\hat{\sigma}^3$. Now, the solutions η_α and $\hat{\eta}_\alpha$ of the respective equations

$$P(T \le \eta_\alpha) = \alpha\,, \quad P(T^* \le \hat{\eta}_\alpha \,|\, \mathcal{X}) = \alpha\,,$$

admit the Cornish-Fisher expansions

$$\begin{aligned} \eta_\alpha &= z_\alpha + n^{-1/2}\, Q_1^{\mathrm{cf}}(z_\alpha) + n^{-1}\, Q_2^{\mathrm{cf}}(z_\alpha) + \dots, \\ \hat{\eta}_\alpha &= z_\alpha + n^{-1/2}\, \widehat{Q}_1^{\mathrm{cf}}(z_\alpha) + n^{-1}\, \widehat{Q}_2^{\mathrm{cf}}(z_\alpha) + \dots . \end{aligned}$$

On subtracting these expansions we deduce that

$$\begin{aligned}\hat{\eta}_\alpha - \eta_\alpha &= n^{-1/2}\,\{\widehat{Q}_1^{\rm cf}(z_\alpha) - Q_1^{\rm cf}(z_\alpha)\} + n^{-1}\,\{\widehat{Q}_2^{\rm cf}(z_\alpha) - Q_2^{\rm cf}(z_\alpha)\} + \ldots \\ &= n^{-1/2}\,\{Q_1(z_\alpha) - \widehat{Q}_1(z_\alpha)\} + O_p(n^{-3/2})\,,\end{aligned}$$

where we have used the fact that $Q_1^{\rm cf} = -Q_1$, $\widehat{Q}_1^{\rm cf} = -\widehat{Q}_1$ and $\widehat{Q}_2^{\rm cf} = Q_2^{\rm cf} + O_p(n^{-1/2})$.

Now,

$$\widehat{Q}_1(u) - Q_1(u) = \tfrac{1}{6}\,(\hat{\gamma} - \gamma)\,\gamma_x\,(1 - u^2)\,, \tag{5}$$

$$\hat{\eta}_\alpha - \eta_\alpha = n^{-1/2}\,\{Q_1(z_\alpha) - \widehat{Q}_1(z_\alpha)\} + O_p(n^{-3/2})\,. \tag{6}$$

It may be proved by Taylor expansion that

$$\begin{aligned}\hat{\gamma} &= \frac{n^{-1}\sum_i \{\epsilon_i - \bar{\epsilon} - (x_i - \bar{x})\,(\hat{d} - d)\}^3}{[n^{-1}\sum_i \{\epsilon_i - \bar{\epsilon} - (x_i - \bar{x})\,(\hat{d} - d)\}^2]^{3/2}} \\ &= \gamma + n^{-1/2}\,U + O_p(n^{-1})\,,\end{aligned} \tag{7}$$

where

$$U = n^{-1/2}\sum_{i=1}^{n}\left\{(\delta_i^3 - \gamma) - \tfrac{3}{2}\,\gamma\,(\delta_i^2 - 1) - 3\,\delta_i\right\}$$

and $\delta_i = \epsilon_i/\sigma$. Combining results (5)–(7) we deduce that

$$\hat{\eta}_\alpha - \eta_\alpha = -n^{-1}\,\tfrac{1}{6}\,U\,\gamma_x\,(1 - z_\alpha^2) + O_p(n^{-3/2})\,.$$

That is,

$$\hat{\eta}_\alpha - \eta_\alpha = -n^{-1}\,c\,U + O_p(n^{-3/2})$$

where $c = \frac{1}{6}\,\gamma_x\,(1 - z_\alpha^2)$. Therefore,

$$\begin{aligned} P(d \in \hat{J}_1) &= P\left\{d < \hat{d} - n^{-1/2}\,(\hat{\sigma}/\sigma_x)\,\hat{\eta}_\alpha\right\} \\ &= P(T > \hat{\eta}_\alpha) \\ &= P\left\{T + n^{-1}\,c\,U > \eta_\alpha + O_p(n^{-3/2})\right\}\,, \\ &= P\left(T + n^{-1}\,c\,U > \eta_\alpha\right) + O(n^{-3/2})\,, \end{aligned}$$

assuming we can treat the "$O_p(n^{-3/2})$" inside the probability as though it were deterministic, and take it outside.

That is:

$$P(d \in \hat{J}_1) = P(T + n^{-1}\,c\,U > \eta_\alpha) + O(n^{-3/2}). \tag{8}$$

It can be proved that, for any choice of the constant c, the first four moments (and hence also the first four cumulants) of $T + n^{-1}\,c\,U$ are identical to those of T, up to but not including terms of order $n^{-3/2}$. Hence, recalling the way in which moments influence Edgeworth expansions,

$$P(T + n^{-1}\,c\,U > \eta_\alpha) = P(T > \eta_\alpha) + O(n^{-3/2}) = 1 - \alpha + O(n^{-3/2})\,.$$

Therefore, by (8),

$$P(d \in \hat{J}_1) = 1 - \alpha + O(n^{-3/2})\,,$$

as had to be proved.

24.16 THE OTHER PERCENTILE-METHOD INTERVAL

Recall that the percentile-method interval $\hat{I}_{12}$ is based on bootstrapping $\hat{d} - d$; that is, it is based on approximating the distribution of this quantity by the conditional distribution of $\hat{d}^* - \hat{d}$.

The "other" percentile method is based on using the conditional distribution of $\hat{d}^*$ to approximate the distribution of $\hat{d}$. It leads to the interval

$$\widehat{K}_1 = \left(-\infty, \hat{d} + n^{-1/2}\,(\hat{\sigma}/\sigma_x)\,\hat{\eta}_{1-\alpha}\right) = (-\infty, \hat{\zeta}_{1-\alpha})\,,$$

where $\hat{\zeta}_{1-\alpha}$ is an approximation to $\zeta_{1-\alpha}$, these two quantities being defined by

$$P(\hat{d}^* \le \hat{\zeta}_{1-\alpha} \,|\, \mathcal{X}) = 1 - \alpha\,, \quad P(\hat{d} \le \zeta_{1-\alpha}) = 1 - \alpha\,.$$

However, $\widehat{K}_1$ has coverage error of size $n^{-1/2}$, not n^{-1}. In particularly, $\widehat{K}_1$ does not enjoy the accuracy of the percentile-method interval $\hat{I}_{12}$. This is a consequence of it addressing the wrong tail of the distribution of $\hat{d}$.

24.17 PROPERTIES OF CONFIDENCE INTERVALS FOR THE CONDITIONAL MEAN, $y_0 = E(Y \,|\, x = x_0)$, AND THE INTERCEPT, c

Recall that $y_0 = c + x_0\,d$, which in turn equals c when $x_0 = 0$. Therefore we can treat confidence intervals for c as a special case of those for y_0.

Noting that $\hat{y}_0 = \hat{c} + x_0\,\hat{d}$, redefine

$$\begin{aligned} S &= n^{1/2}\,(\hat{y}_0 - y_0)/(\sigma\,\sigma_y)\,, \\ T &= n^{1/2}\,(\hat{y}_0 - y_0)/(\hat{\sigma}\,\sigma_y)\,; \end{aligned}$$

and taking $\hat{y}_0^* = \hat{c}^* + x_0\,\hat{d}^*$, redefine

$$\begin{aligned} S^* &= n^{1/2}\,(\hat{y}_0^* - \hat{y}_0)/(\hat{\sigma}\,\sigma_y)\,, \\ T^* &= n^{1/2}\,(\hat{y}_0^* - \hat{y}_0)/(\hat{\sigma}^*\,\sigma_y)\,. \end{aligned}$$

Percentile-method and percentile-t confidence intervals for y_0 and c are given respectively by

$$\begin{aligned} \hat{I}_{12} &= \left(-\infty, \hat{y}_0 - n^{-1/2}\,(\hat{\sigma}/\sigma_y)\,\hat{\xi}_\alpha\right)\,, \\ \hat{J}_1 &= \left(-\infty, \hat{y}_0 - n^{-1/2}\,(\hat{\sigma}/\sigma_y)\,\hat{\eta}_\alpha\right)\,, \end{aligned}$$

where $\sigma_y^2 = 1 + \sigma_x^{-2}\,(x_0 - \bar{x})^2$ and we define $\hat{\xi}_\alpha$ and $\hat{\eta}_\alpha$ by

$$P(S^* \le \hat{\xi}_\alpha \,|\, \mathcal{X}) = \alpha\,, \quad P(T^* \le \hat{\eta}_\alpha \,|\, \mathcal{X}) = \alpha\,,$$

for the new versions of S^* and T^*.

The intervals $\hat{I}_{12}$ and $\hat{J}_1$ have coverage errors $O(n^{-1/2})$ and $O(n^{-1})$, respectively. These results, unlike those for slope, are conventional.

25 BOOTSTRAP METHODS FOR NONPARAMETRIC CURVE ESTIMATION

25.1 POINTWISE VERSUS SIMULTANEOUS CONFIDENCE REGIONS

We shall dispose of this topic first, so that we can focus subsequently on other issues.

Suppose we have an estimator $\hat{g}$ of a function g on an interval $\mathcal{I}$, and, for a given level $1-\alpha$ of probability, have constructed a confidence region, or "tube," for g, consisting of a boundary above and a boundary below the curve represented by the formula $y = \hat{g}(x)$, for $x \in \mathcal{I}$.

The region can be interpreted as the union of intervals $(\hat{g}_1(x), \hat{g}_2(x))$, for $x \in \mathcal{I}$. Of course, $\hat{g}_1$ and $\hat{g}_2$ are constructed from data, and satisfy $\hat{g}_1 \le \hat{g}_2$.

Such a region is commonly referred to a "$(1-\alpha)$-level confidence region for g on the interval $\mathcal{I}$."

We can interpret the statement in two ways. Either (i) the interval $(\hat{g}_1(x), \hat{g}_2(x))$ covers $g(x)$ with probability approximately $1-\alpha$, for each $x \in \mathcal{I}$; or (ii) the probability that the graph represented by the equation $y = g(x)$ lies within the tube, converges to $1-\alpha$ as n increases.

Interpretations (i) and (ii) are generally referred to as "pointwise" and "simultaneous," respectively.

In conventional parametric problems the pointwise interpretation seems generally to be favoured. For example, in regression we often wish to "predict" the value of $E(Y \mid X = x_0)$, denoting the regression mean, for only a small number of values of x_0.

However, taking the simultaneous interpretation causes no difficulty in the parametric case. Bootstrap methods are just as easily pressed into use there as in the pointwise context. In particular, in parametric problems, both pointwise and simultaneous confidence regions are of width $n^{-1/2}$.

This close relationship vanishes in nonparametric cases, however. There, simultaneous confidence regions are generally an order of magnitude wider than their parametric counterparts.

Although the factor by which the width increases is proportional only to $(\log n)^{1/2}$, in asymptotic terms, the increase is generally substantial, and this alone causes simultaneous bands to be unpopular.

When coupled with the relative lack of interest in predicting the value of $E(Y \mid X = x_0)$ simultaneously for many values of x_0, this means that the pointwise interpretation is the obvious choice in at least the setting of nonparametric regression. We shall adopt it in the density estimation context, too.

Our treatment of confidence regions in the setting of nonparametric curve estimation will address only the case of nonparametric density estimation. Nonparametric regres-