

EXAMPLE OF BOOTSTRAP INCONSISTENCY

Distribution of the Square of the Sample Average

Let $\{X_i; i = 1, \dots, n\}$ be a random sample from the $N(\mathbf{m}, 1)$ distribution. Let \bar{X} denote the sample average. Let F_n be the EDF of the sample. Set $T_n = n^{1/2}(\bar{X}^2 - \mathbf{m}^2)$ if $\mathbf{m} \neq 0$ and $T_n = n\bar{X}^2$ otherwise.

Asymptotic Distribution of T_n :

Case 1: $\mathbf{m} \neq 0$. Write

$$\begin{aligned} n^{1/2}(\bar{X}^2 - \mathbf{m}^2) &= n^{1/2} \{[(\bar{X} - \mathbf{m}) + \mathbf{m}]^2 - \mathbf{m}^2\} \\ &= n^{1/2} \mathbf{m}^2 \{[1 + (\bar{X} - \mathbf{m}) / \mathbf{m}]^2 - 1\} \\ &= n^{1/2} \mathbf{m}^2 [2(\bar{X} - \mathbf{m}) / \mathbf{m} + O_p(n^{-1})] \\ &= 2\mathbf{m}\bar{Z} + o_p(1), \end{aligned}$$

where $Z \sim N(0,1)$. So $T_n = 2\mathbf{m}N(0,1)$ if $\mathbf{m} \neq 0$.

Case 2: $\mathbf{m} = 0$. In this case $n^{1/2}\bar{X} \sim N(0,1)$, so $T_n = n\bar{X}^2 \sim \mathbf{c}^2(1)$ (chi-square with one degree of freedom).

The Bootstrap

Implement the bootstrap by sampling the EDF of the data (i.e., sampling the estimation data randomly with replacement). Denote the bootstrap sample by $\{X_i^*; i = 1, \dots, n\}$. The bootstrap analog of T_n is $T_n^* = n^a[(\bar{X}^*)^2 - \bar{X}^2]$, where $a = 1/2$ if $\mathbf{m} \neq 0$ and $a = 1$ otherwise.

Case 1: $\mathbf{m} \neq 0$

$$\begin{aligned} T_n^* &= n^{1/2}[(\bar{X}^*)^2 - \bar{X}^2] \\ &= n^{1/2} \left\{ n^{-1} \sum_{i=1}^n (X_i^* - \bar{X}) + \bar{X} \right\}^2 - \bar{X}^2. \end{aligned}$$

$$= n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X}) \Big| n^{-1} \sum_{i=1}^n (X_i^* - \bar{X}) + 2\bar{X} \Big|.$$

But $\bar{X} \rightarrow^{a.s.} \mathbf{m}$. Moreover by Mammen's theorem,

$$n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X}) \rightarrow^d N(0,1).$$

Therefore,

$$n^{-1} \sum_{i=1}^n (X_i^* - \bar{X}) \rightarrow^p 0$$

Thus, combining results gives $T_n^* \rightarrow^d 2\mathbf{m}N(0,1)$ almost surely. The bootstrap is consistent. ("Almost surely" refers to the fact that the distribution of T_n^* is conditional on the data.)

Case 2: $\mathbf{m}=0$: By the same algebra as before,

$$T_n^* = n[(\bar{X}^*)^2 - \bar{X}^2]$$

$$= n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X}) \Big| n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X}) + 2n^{1/2} \bar{X} \Big|.$$

Mammen's theorem still gives

$$n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X}) \rightarrow^d N(0,1)$$

Therefore, conditional on the data

$$T_n^* \rightarrow^d Z^2 + 2\mathbf{w}Z,$$

where $Z \sim N(0,1)$ and \mathbf{w} is a constant that varies randomly among estimation samples.

Conclusion: The bootstrap is inconsistent when $\mathbf{m}=0$ unless the bootstrap sample is drawn from a distribution whose sample average is 0.

EDGEWORTH EXPANSIONS

Let $\{X_i: i = 1, \dots, n\}$ be realizations of a random variable X that has mean \mathbf{m} and standard deviation \mathbf{s} . Let Z_i denote the standardized form of X_i :

$$Z_i = \frac{X_i - \mathbf{m}}{\mathbf{s}}.$$

Define

$$S_n = n^{-1/2} \sum_{i=1}^n Z_i.$$

The problem is to find the probability distribution or density of S_n .

If we know the probability distribution of X , then we can calculate the probability distribution of S_n . Suppose we do not know the distribution of X . What can be done to get an approximation to the distribution of S_n ? One possibility is to use asymptotic distribution theory. Specifically, we know from the Lindeberg-Levy theorem that under regularity conditions,

$$P(S_n \leq \mathbf{x}) \rightarrow \Phi(\mathbf{x})$$

as $n \rightarrow \infty$. So if n is sufficiently large, we can approximate the CDF or density of S_n arbitrarily well with the normal distribution or density.

But this approximation may not be very accurate if n is small. One consequence can be that the empirical probability that a t test rejects a correct hypothesis about \mathbf{m} may be much different from the nominal rejection probability.

Therefore, we seek an approximation to the distribution of S_n that is likely to be more accurate in small samples. One way is to try to find an approximation of the form

$$p_n(z) = \mathbf{f}(z) + n^{-1/2} g_1(z) + n^{-1} g_2(z) + \dots,$$

where p_n is the probability density function of z . In very large samples, the higher-order terms are negligible, but this is not necessarily so in small samples.

To derive the expansion, let \mathbf{y} denote the characteristic function of Z :

$$\mathbf{y}_Z(t) = \int e^{i t z} f(z) dz,$$

where f is the density of Z and $i = \sqrt{-1}$. The characteristic function of S_n is

$$\begin{aligned}
y_n(t) &= E \exp(it S_n) \\
&= E \exp\left[i \frac{t}{n^{1/2}} \sum_{j=1}^n Z_j \right] \\
&= E \prod_{j=1}^n \exp\left[i \frac{t}{n^{1/2}} Z_j \right] \\
&= \prod_{j=1}^n E \exp\left[i \frac{t}{n^{1/2}} Z_j \right] \\
&= \left[E \exp\left[i \frac{t}{n^{1/2}} Z \right] \right]^n.
\end{aligned}$$

Using the inversion formula for characteristic functions, we get

$$\begin{aligned}
p_n(z) &= \frac{1}{2\pi} \int e^{-itz} [y_Z(t/n^{1/2})]^n dt \\
&= \frac{1}{2\pi} \int \exp\left[-itz + n \log y_Z\left(\frac{t}{n^{1/2}}\right) \right] dt.
\end{aligned}$$

Now make a Taylor series expansion of $\log y_Z(t)$ about $t = 0$:

$$h(t) = \log y_Z(t)$$

$$h(0) = \log y_Z(0) = 0$$

$$h'(t) = \frac{y_Z'(t)}{y_Z(t)}$$

$$h''(t) = -\frac{y_Z'(t)^2}{y_Z(t)^2} + \frac{y_Z''(t)}{y_Z(t)}$$

$$h'''(t) = \frac{2y_Z'(t)^3}{y_Z(t)^3} - \frac{3y_Z'(t)y_Z''(t)}{y_Z(t)^2} + \frac{y_Z'''(t)}{y_Z(t)}.$$

We could continue with higher-order terms, but this is not necessary for now.

Now see that

$$h''(0) = y_Z''(0) = -s_Z^2 = -1.$$

Also

$$\begin{aligned}
h'''(0) &= \mathbf{y}_Z'''(0) = \frac{d^3}{dt^3} \int e^{itz} f(z) dz \\
&= -\mathbf{i} \int z^3 e^{itz} f(z) dz.
\end{aligned}$$

Therefore,

$$\mathbf{y}_Z'''(0) = -\mathbf{i} \int z^3 f(z) dz = -\mathbf{i} \mathbf{m}_3.$$

The Taylor series expansion is now

$$\log \mathbf{y}_Z(t) = -\frac{1}{2} t^2 - \frac{\mathbf{i}}{6} \mathbf{m}_3 t^3 + O(t^4)$$

$$\log \mathbf{y}_Z \left(\frac{t}{n^{1/2}} \right) = -\frac{1}{2} \left(\frac{t}{n^{1/2}} \right)^2 - \frac{\mathbf{i}}{6} \mathbf{m}_3 \left(\frac{t}{n^{1/2}} \right)^3 + O \left(\frac{1}{n^2} \right)$$

$$n \log \mathbf{y}_Z \left(\frac{t}{n^{1/2}} \right) = -\frac{t^2}{2} - \frac{\mathbf{i}}{6} \mathbf{m}_3 \left(\frac{t^3}{n^{1/2}} \right) + O \left(\frac{1}{n} \right).$$

Substituting this into the equation for $p_n(z)$ gives

$$\begin{aligned}
p_n(z) &= \frac{1}{2\mathbf{p}} \int \exp \left[-itz - \frac{1}{2} t^2 - \frac{\mathbf{i}}{6} \mathbf{m}_3 \left(\frac{t^3}{n^{1/2}} \right) + O \left(\frac{1}{n} \right) \right] dt \\
&= \frac{1}{2\mathbf{p}} \int e^{-itz - (1/2)t^2} \left[1 - \frac{\mathbf{i}}{6} \mathbf{m}_3 \left(\frac{t^3}{n^{1/2}} \right) + O \left(\frac{1}{n} \right) \right] dt.
\end{aligned}$$

Recall that $e^{-(1/2)t^2}$ is the characteristic function of the standard normal distribution. Moreover, under the Cramér condition, the $O(1/n)$ remainder term can be taken outside the integral. Therefore,

$$p_n(z) = \mathbf{f}(z) - \frac{\mathbf{i}}{6} \frac{\mathbf{m}_3}{n^{1/2}} \mathbf{m}_3 \frac{1}{2\mathbf{p}} \int t^3 e^{-itz - (1/2)t^2} dt + O \left(\frac{1}{n} \right).$$

So the \mathbf{m}_3 term is the desired $O(n^{-1/2})$ term of the expansion.

To carry out the integral, note that

$$\begin{aligned}
\int t^3 e^{-itz - (1/2)t^2} dt &= \mathbf{i} \frac{d^3}{dz^3} \int e^{-itz - (1/2)t^2} dt \\
&= -\mathbf{i} \frac{d^3}{dz^3} \mathbf{f}(z) \\
&= \mathbf{i}(3z - z^3) \mathbf{f}(z).
\end{aligned}$$

Therefore,

$$p_n(z) = f(z) + \frac{1}{6} \frac{m_3}{n^{1/2}} (z^3 - 3z) + O(n^{-1}).$$

Thus, we have an approximation through $O(n^{-1/2})$ to the density of S_n .

If we had kept the 4th derivative terms in the Taylor series expansion, these would have produced an $O(n^{-1})$ term in the expansion of p_n .

It turns out that this expansion idea can be applied to a very large class of functions of random variables, not just sums.

Reference: Bhattacharya, R.N. and J.K. Ghosh (1978). On the Validity of the Formal Edgeworth Expansion, *Annals of Statistics*, 6, 434-451.

Any smooth function $G(\bar{Z})$ of sample means that satisfies certain regularity conditions has an expansion of the form

$$P\left\{\frac{G(\bar{Z}) - G(\mathbf{m})}{\mathbf{s}_G} \leq z\right\} = \Phi(z) + n^{-1/2} g_1(z) + n^{-1} g_2(z) + \dots + n^{-j/2} g_j(z) + O(n^{-(j+1)/2}).$$

Note that this is an *asymptotic expansion*. It does not necessarily converge as an infinite series.

Analytic Edgeworth expansions usually cannot be used for inference because they are very hard to calculate. This is where the bootstrap comes in.

ERROR IN BOOTSTRAP REJECTION PROBABILITY OF A SYMMETRICAL TEST

These notes provide an informal derivation of equation (3.27) of the *Handbook* chapter. This equation is

$$(3.27) \quad P(|T_n| > z_{n,\mathbf{a}/2}^*) = \mathbf{a} + O(n^{-2}).$$

Start with the Cornish-Fisher inversions

$$(3.23) \quad z_{n,\mathbf{a}/2} = z_{\infty,\mathbf{a}/2} - \frac{1}{n} \frac{g_2(z_{\infty,\mathbf{a}/2}, F_0)}{\mathbf{f}(z_{\infty,\mathbf{a}/2})} + O(n^{-2}),$$

where \mathbf{f} is the standard normal density function, and

$$(3.24) \quad z_{n,\mathbf{a}/2}^* = z_{\infty,\mathbf{a}/2} - \frac{1}{n} \frac{g_2(z_{\infty,\mathbf{a}/2}, F_n)}{\mathbf{f}(z_{\infty,\mathbf{a}/2})} + O(n^{-2})$$

almost surely. Proceed to equation (A.3) of the *Handbook* chapter. Observe that the cumulants of $T_n - n^{-3/2}n^{1/2}r_3(\bar{Z})$ and $T_n + n^{-3/2}n^{1/2}r_3(\bar{Z})$ differ by at most $O(n^{-3/2})$. Since the cumulants enter Edgeworth terms whose sizes are at most $O(n^{-1/2})$, the differences are negligible. That is, with an error of at most $O(n^{-2})$, the cumulants of $T_n - n^{-3/2}n^{1/2}r_3(\bar{Z})$ and $T_n + n^{-3/2}n^{1/2}r_3(\bar{Z})$ are the same.

Now denote the vectors of cumulants by \mathbf{k}_n . From equation (3.9) of the *Handbook* chapter, get the following Edgeworth expansions:

$$\begin{aligned} & P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2})] \\ &= \Phi[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2})] + \frac{1}{n^{1/2}} g_1[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2}), \mathbf{k}_n] \\ &+ \frac{1}{n} g_2[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2}), \mathbf{k}_n] + \frac{1}{n^{3/2}} g_3[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2})] + O(n^{-2}) \end{aligned}$$

By a Taylor series expansion

$$\begin{aligned} & P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2})] \\ &= \Phi[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2})] + \frac{1}{n^{1/2}} g_1[z_{\infty,\mathbf{a}/2} + n^{-1}r_2(z_{\infty,\mathbf{a}/2}), \mathbf{k}_n] \\ &+ \frac{1}{n} g_2(z_{\infty,\mathbf{a}/2}, \mathbf{k}_n) + \frac{1}{n^{3/2}} g_3(z_{\infty,\mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2}) \end{aligned}$$

Similarly,

$$\begin{aligned}
& P[T_n + n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq -z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&= \Phi[-z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] + \frac{1}{n^{1/2}}g_1[-z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2}), \mathbf{k}_n] \\
&+ \frac{1}{n}g_2(-z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + \frac{1}{n^{3/2}}g_3(-z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2}).
\end{aligned}$$

Recall that g_1 and g_3 are even functions and that g_2 is an odd function. Therefore

$$\begin{aligned}
& P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&- P[T_n + n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq -z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&= \Phi[z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] - \Phi[-z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&+ \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2}) \\
&= 2\Phi[z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] - 1 \\
&+ \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2})
\end{aligned}$$

A Taylor series expansion gives

$$\begin{aligned}
& P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&- P[T_n + n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq -z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
&= 2\Phi(z_{\infty, \mathbf{a}/2}) - 1 + \frac{2}{n}r_2(z_{\infty, \mathbf{a}/2})\mathbf{f}(z_{\infty, \mathbf{a}/2}) \\
&+ \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2}) \\
&= 1 - \mathbf{a} + \frac{2}{n}r_2(z_{\infty, \mathbf{a}/2})\mathbf{f}(z_{\infty, \mathbf{a}/2}) \\
&+ \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2})
\end{aligned}$$

Recall from (A.4) of the *Handbook* chapter that

$$r_2(z_{\infty, \mathbf{a}/2}) = -\frac{g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k})}{\mathbf{f}(z_{\infty, \mathbf{a}/2})},$$

where \mathbf{k} is a vector of population cumulants that is conformable with \mathbf{k}_n . Therefore

$$\begin{aligned}
& P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
& - P[T_n + n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq -z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
& = 1 - \mathbf{a} - \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}) + \frac{2}{n}g_2(z_{\infty, \mathbf{a}/2}, \mathbf{k}_n) + O(n^{-2}).
\end{aligned}$$

It turns out that $\mathbf{k} - \mathbf{k}_n = o(n^{-1})$. Therefore,

$$\begin{aligned}
& P[T_n - n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq z_{\infty, \mathbf{a}/2} + n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
& - P[T_n + n^{-3/2}n^{1/2}r_3(\bar{Z}) \leq -z_{\infty, \mathbf{a}/2} - n^{-1}r_2(z_{\infty, \mathbf{a}/2})] \\
& = 1 - \mathbf{a} + O(n^{-2}),
\end{aligned}$$

which is the desired result.

WHY A CORRECTION FACTOR IS NEEDED IN THE BLOCK BOOTSTRAP t STATISTIC

Consider the t statistic for testing the hypothesis $H_0: \mathbf{q} = 0$ based on the moment condition $E(X - \mathbf{q}) = 0$. In other words, we test the hypothesis that the population mean is zero. The test statistic is

$$(1) \quad t = \frac{n^{1/2} \bar{X}}{V_n^{1/2}},$$

where \bar{X} is the sample average, and V_n is an estimator of $\text{Var}(n^{1/2} \bar{X})$. Suppose that $\{X_i\}$ is an uncorrelated sequence. (Later I will assume that it is iid.) Then

$$\begin{aligned} V_n &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{1}{n} (n^{1/2} \bar{X})^2. \end{aligned}$$

Let $\mathbf{s}^2 = E(X_i^2)$. Then a Taylor-series expansion gives

$$V_n^{-1/2} = (\mathbf{s}^2)^{-1/2} \left[-\frac{V_n - \mathbf{s}^2}{2\mathbf{s}^2} + \frac{3}{8} \frac{(V_n - \mathbf{s}^2)^2}{\mathbf{s}^2} + \dots \right],$$

so

$$t = \frac{n^{1/2} \bar{X}}{\mathbf{s}} \left[-\frac{V_n - \mathbf{s}^2}{2\mathbf{s}^2} + \frac{3}{8} \frac{(V_n - \mathbf{s}^2)^2}{\mathbf{s}^2} + \dots \right].$$

Now observe that

$$n^{1/2} (V_n - \mathbf{s}^2) = \frac{1}{n^{1/2}} \sum_{i=1}^n (X_i^2 - \mathbf{s}^2) - \frac{1}{n^{1/2}} (n^{1/2} \bar{X})^2$$

and

$$[n^{1/2} (V_n - \mathbf{s}^2)]^2 = \frac{1}{n} \sum_{i=1}^n (X_i^2 - \mathbf{s}^2)^2 + o_p(n^{-1}).$$

Define

$$b_n = \frac{1}{\mathbf{s}^2 n^{1/2}} \sum_{i=1}^n (X_i^2 - \mathbf{s}^2)^2$$

and

$$c_n = -\frac{1}{s^2 n^{1/2}} (n^{1/2} \bar{X})^2.$$

Then

$$(2) \quad t = \frac{n^{1/2} \bar{X}}{s} \left\| -\frac{b_n + c_n}{2n^{1/2}} + \frac{3b_n^2}{8n} \right\| + o_p(n^{-1}).$$

Observe that

$$\text{Var} \left\| \frac{n^{1/2} \bar{X}}{s} \right\| = 1$$

exactly.

A block bootstrap analog of t might be formed by replacing all variables in (1) with bootstrap analogs. This gives

$$\tilde{t} = \frac{n^{1/2} (\bar{X}^* - \bar{X})}{(V_n^*)^{1/2}},$$

where the block bootstrap sample is $\{X_i^*: i = 1, \dots, n\}$,

$$\bar{X}^* = \frac{1}{n} \sum_{i=1}^n X_i^*,$$

and

$$V_n^* = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2.$$

Taylor series arguments identical to those used with t now give

$$(3) \quad \tilde{t} = \frac{n^{1/2} \bar{X}^*}{s_n} \left\| -\frac{b_n^* + c_n^*}{2n^{1/2}} + \frac{3(b_n^*)^2}{8n} \right\| + o_p(n^{-1}),$$

where

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

$$b_n^* = \frac{1}{s_n^2 n^{1/2}} \sum_{i=1}^n [(X_i^*)^2 - s_n^2]^2,$$

and

$$c_n^* = -\frac{1}{s_n^2 n^{1/2}} (n^{1/2} \bar{X}^*)^2.$$

Now t and t^* have the same algebraic forms, so it might seem that they have the same Edgeworth expansions apart from replacing population cumulants with sample cumulants in

the bootstrap expansion. But this is not correct. The reason is that with the block bootstrap with b blocks of length ℓ ,

$$\begin{aligned} \text{Var}^*(n^{1/2}\bar{X}^*) &= \frac{1}{n} \sum_{i=0}^{b-1} \sum_{j=1}^{\ell} \sum_{k=1}^{\ell} (X_{i\ell+j} - \bar{X})(X_{i\ell+k} - \bar{X}) \\ &\equiv \tilde{s}_n^2. \end{aligned}$$

Therefore,

$$\text{Var}^* \left(\frac{n^{1/2}\bar{X}^*}{s_n} \right) \neq 1.$$

Consequently, (2) and (3) do not have identical Edgeworth expansions apart from replacing population with bootstrap cumulants.

The correction factor is designed to solve this problem. Its purpose is to make the leading term of the right-hand side of (3) have variance 1 and to do this without introducing new (bootstrap) stochastic terms that would affect the structure of the Edgeworth expansion. The correction factor is

$$\mathbf{t}_n = (s_n^2 / \tilde{s}_n^2)^{1/2}.$$

Observe that this is non-stochastic relative to the probability measure induced by bootstrap sampling. The corrected bootstrap t statistic is

$$t^* = \mathbf{t}_n \tilde{t}$$

$$= \frac{n^{1/2}\bar{X}^*}{\tilde{s}_n} \left[-\frac{b_n^* + c_n^*}{2n^{1/2}} + \frac{3(b_n^*)^2}{8n} \right] + o_{P^*}(n^{-1}).$$

The leading term of t^* has exact bootstrap variance 1. In other respects, it has the same structure as t , so t^* and t have Edgeworth expansions that differ only by replacing population cumulants with sample cumulants.

What happens if the correction factor is omitted? To see this, let \mathbf{k}_n denote the set of sample cumulants. Then the bootstrap Edgeworth expansion of t^* is

$$(4) \quad P^*(t^* \leq z) = \Phi(z) + n^{-1/2} g_1(z, \mathbf{k}_n) + n^{-2} g_2(z, \mathbf{k}_n) + o(n^{-1})$$

almost surely. But

$$P^*(\tilde{t} \leq z) = P^*(t^* \leq \mathbf{t}_n z)$$

$$= \Phi(\mathbf{t}_n z) + n^{-1/2} g_1(\mathbf{t}_n z, \mathbf{k}_n) + n^{-2} g_2(\mathbf{t}_n z, \mathbf{k}_n) + o(n^{-1})$$

almost surely. A Taylor series expansion about $\mathbf{t}_n = 1$ yields

$$\begin{aligned}
P^*(t^* \leq z) &= \Phi(z) + z\mathbf{f}(z)(\mathbf{t}_n - 1) - (1/2)z^3\mathbf{f}(z)(\mathbf{t}_n - 1)^2 \\
&\quad + n^{-1/2}g_1(z, \mathbf{k}_n) + n^{-1/2}g_1'(z, \mathbf{k}_n)(\mathbf{t}_n - 1) + n^{-1}g_2(z, \mathbf{k}_n) \\
&\quad + O[(\mathbf{t}_n - 1)^3] + O[n^{-1/2}(\mathbf{t}_n - 1)^2] + O[n^{-1}(\mathbf{t}_n - 1)] + o(n^{-1})
\end{aligned}$$

uniformly over z . For the case of a symmetrical test, this simplifies to

$$\begin{aligned}
P^*(|\tilde{t}| \leq z) &= 2\Phi(z) - 1 + 2z\mathbf{f}(z)(\mathbf{t}_n - 1) - z^3\mathbf{f}(z)(\mathbf{t}_n - 1)^2 \\
&\quad - 2n^{-1}g_2(z, \mathbf{k}_n) + O[(\mathbf{t}_n - 1)^3] + o(n^{-1}).
\end{aligned}$$

The population expansion does not have the terms involving $(\mathbf{t}_n - 1)$ or $(\mathbf{t}_n - 1)^2$. To provide a refinement through $O(n^{-1})$, these terms must be $o_p(n^{-1})$. That is, we need

$$(5) \quad \mathbf{t}_n - 1 = o_p(n^{-1}).$$

I will now show that (5) does not hold in general. To do this most simply, let $\{X_i\}$ be a sequence of iid random variables with means of zero. Because $s_n^2 \rightarrow E(X^2)$, $\tilde{s}_n^2 \rightarrow E(X^2)$, and $E(X^2) \neq 0$, the rate at which \mathbf{t}_n converges to zero is the same as the rate at which $\tilde{s}_n - s_n$ converges to zero. But

$$\begin{aligned}
\tilde{s}_n^2 - s_n^2 &= \frac{1}{n} \sum_{i=0}^{b-1} \sum_{j=1}^{\ell} \sum_{\substack{k=1 \\ k \neq j}}^{\ell} (X_{i\ell+j} - \bar{X})(X_{i\ell+k} - \bar{X}) \\
&= \frac{1}{n} \sum_{i=0}^{b-1} \sum_{j=1}^{\ell} \sum_{\substack{k=1 \\ k \neq j}}^{\ell} X_{i\ell+j} X_{i\ell+k} - (\ell-1)\bar{X}^2.
\end{aligned}$$

Therefore,

$$E(\tilde{s}_n^2 - s_n^2) = \frac{\ell-1}{n} E(X^2).$$

Therefore $\mathbf{t}_n - 1$ has size at least $O_p(\ell/n)$, and $\mathbf{t}_n \neq o_p(n^{-1})$. Therefore, the block bootstrap does not provide an $O(n^{-1})$ asymptotic refinement for the uncorrected symmetrical t test.