The Annals of Statistics
2001, Vol. 29, No. 5, 1443–1468

BOOTSTRAPPING NONPARAMETRIC DENSITY ESTIMATORS WITH EMPIRICALLY CHOSEN BANDWIDTHS

BY PETER HALL AND KEE-HOON KANG

Australian National University and Australian National University and Hankuk University of Foreign Studies

We examine the way in which empirical bandwidth choice affects distributional properties of nonparametric density estimators. Two bandwidth selection methods are considered in detail: local and global plug-in rules. Particular attention is focussed on whether the accuracy of distributional bootstrap approximations is appreciably influenced by using the resample version $\hat{h}^*$, rather than the sample version $\hat{h}$, of an empirical bandwidth. It is shown theoretically that, in marked contrast to similar problems in more familiar settings, no general first-order theoretical improvement can be expected when using the resampling version. In the case of local plug-in rules, the inability of the bootstrap to accurately reflect biases of the components used to construct the bandwidth selector means that the bootstrap distribution of $\hat{h}^*$ is unable to capture some of the main properties of the distribution of $\hat{h}$. If the second derivative component is slightly undersmoothed then some improvements are possible through using $\hat{h}^*$, but they would be difficult to achieve in practice. On the other hand, for global plug-in methods, both $\hat{h}$ and $\hat{h}^*$ are such good approximations to an optimal, deterministic bandwidth that the variations of either can be largely ignored, at least at a first-order level. Thus, for quite different reasons in the two cases, the computational burden of varying an empirical bandwidth across resamples is difficult to justify.

1. Introduction. One interpretation of the manner in which bootstrap methods work is that they model the relationship between the population and the sample by that between the sample and a "resample," drawn from the sample by sampling randomly with replacement. A tenet of this viewpoint is that, when constructing the bootstrap version of the population-sample relationship, each part of a "statistic" (we use the term in a general sense) that depends on the population should be replaced by its sample version, and each part that depends on the sample should be replaced by its resample counterpart. If this nexus is broken then the quality of the approximation is likely to be degraded.

Take, for example, the problem of estimating the distribution of the statistic $T = (\overline{X} - m)/S$, where m denotes the population mean, $\overline{X} = n^{-1}\sum_i X_i$ is the sample mean, and $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$ is the sample variance, with both $\overline{X}$ and S^2 being computed from the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$. Taking $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ to be a resample drawn by sampling randomly,

Received February 2000; revised February 2001.

AMS 2000 *subject classifications*. Primary 62G15; secondary 62G20.

Key words and phrases. Bootstrap methods, confidence interval, Edgeworth expansion, kernel methods, nonparametric estimation, plug-in rules, rate of convergence, second-order accuracy, smoothing parameter.

with replacement, from $\mathcal{X}$, we define $\overline{X}^*$ and $(S^*)^2$ to be its mean and variance, respectively. Percentile-t bootstrap methods [see, e.g., Hall (1992a), pages 15–16; Efron and Tibshirani (1993), pages 160–167; Davison and Hinkley (1997), pages 29–30] suggest using the distribution of $T^* = (\overline{X}^* - \overline{X})/S^*$, conditional on $\mathcal{X}$, as an approximation to the distribution of T. The approximation is second-order accurate, in the sense that the two distributions are within $O_p(n^{-1})$, not just $O_p(n^{-1/2})$, of one another. However, if we replace S^* by S in the definition of T^* then this property is lost, and the accuracy of the approximation is only $O_p(n^{-1/2})$.

The statistic T depends on a population quantity, m, and on the sample quantities $\overline{X}$ and S; and these should be replaced by the sample quantity $\overline{X}$ and resample quantities $\overline{X}^*$ and S^*, respectively, if performance is not to be sacrificed. In this article we shall address a closely related problem in the context of density estimation. The problem often arises when using the bootstrap to estimate the variability of a density estimator, for example in terms of confidence bands. We shall show that, for density estimation (and for curve estimation more generally), the population-sample relationship is much more complex than in more classical circumstances, and that in important, specific cases there is little to be gained by adhering rigorously to the relationship when developing a bootstrap approximation.

One of the problems we address is the following. Consider estimating a probability density, $f(x)$, using a kernel estimator, $\hat{f}(x) = \hat{f}(x|h)$, determined by a bandwidth h and computed from a data set $\mathcal{X}$. In practice, h too would be computed from $\mathcal{X}$; in that case we denote it by $\hat{h}$. Let h_0 denote the theoretically optimal bandwidth that $\hat{h}$ is attempting to capture. When using $\hat{f}(x)$ in connection with bootstrap methods, for example with the aim of constructing a confidence region for $E\{\hat{f}(x|h_0)\}$ or $E\{\hat{f}(x|\hat{h})\}$, we would compute the estimator $\hat{f}$ from a resample $\mathcal{X}^*$ rather than from $\mathcal{X}$. Denote the resulting estimator by $\hat{f}^*(x|h)$. Experience in more conventional settings, such as the percentile-t problem discussed earlier, suggests that $\hat{f}^*$ should be computed using the bootstrap version, $h = \hat{h}^*$, say, of $\hat{h}$. (The bandwidth $\hat{h}^*$ is the same function of $\mathcal{X}^*$ as $\hat{h}$ was of $\mathcal{X}$.) In particular, we should calculate the bootstrap distribution of $\hat{f}^*(x|\hat{h}^*)$ rather than that of $\hat{f}^*(x|\hat{h})$.

Analogously to the percentile-t case, we would expect this approach to reduce the order of error associated with the bootstrap approximation to a distribution. However, we shall show that in several important respects this is false. In particular, let us suppose $\hat{h}$ is a standard local plug-in bandwidth selector for a second-order kernel density estimator $\hat{f}$, based on pilot estimators of f and f''; see, for example, Wand and Jones [(1995), page 41] for the theoretical version of this bandwidth, denoted here by h_0. Then, replacing $\hat{h}$ by $\hat{h}^*$ when computing the bootstrapped estimator $\hat{f}^*(x|\hat{h})$ does not necessarily improve the order of accuracy of a confidence procedure based on the bootstrap. It will of course lead to numerical differences, but unlike the percentile-t case, it will not usually improve performance by an order of magnitude.

The reasons are complex, but result primarily from the fact that standard bootstrap methods are unable to accurately estimate the bias of a curve esti-

mator. Bias terms contribute significantly to the distribution of $\hat{f}(x|\hat{h})$, where $\hat{h}$ is a local bandwidth selector. Another reason for the disparity between the present context and that of the Studentized mean is that replacing $\hat{h}$ by $\hat{h}^*$ does not reproduce the stabilization or pivoting that is effected by Studentizing.

The situation is quite different when bandwidth is chosen using a global plug-in rule. The relative accuracy with which a global bandwidth can be estimated means that the first term in an Edgeworth expansion of the distribution of $\hat{f}(x|h_0)$ is shared by the analogous expansion of the distribution of $\hat{f}(x|\hat{h})$. (On this occasion the bandwidth h_0 is the theoretical version of the global plug-in bandwidth, and $\hat{h}$ is its empirical counterpart.) As a result, the bootstrap distributions of both $\hat{f}^*(x|\hat{h}) - \hat{f}(x|\hat{h})$ and $\hat{f}^*(x|\hat{h}^*) - \hat{f}(x|\hat{h})$ successfully capture the dominant term causing departure of the distributions of both $\hat{f}(x|h_0) - E\{\hat{f}(x|h_0)\}$ and $\hat{f}(x|\hat{h}) - E\{\hat{f}(x|h_0)\}$ from Normality. In particular, changing $\hat{h}$ to $\hat{h}^*$ at the bootstrap step has no impact on first-order accuracy. Moreover, changing the definition of location from $E\{\hat{f}(x|h_0)\}$ to $E\{\hat{f}(x|\hat{h})\}$ in the target distribution has no effect on the first-order term.

These two classes of results, about local and global plug-in bandwidth selectors, respectively, demonstrate a marked dichotomy of properties. Nevertheless, both directly contradict traditional wisdom for bootstrap methods. In both cases, using $\hat{h}^*$ rather than $\hat{h}$ has little theoretical effect on accuracy, in terms of the first order of departure from the asymptotic distribution, of the bootstrap approximation to the distribution of a density estimator. Similar results may be obtained for other plug-in bandwidth selectors; they tend to be either defeated by failure of the bootstrap to capture bias, at least in the case of the usual prescriptions for their pilot bandwidths, or so accurate that resampling the empirical bandwidth selector is largely unnecessary. Scott [(1992), pages 172–177] discusses a range of plug-in rules, including alternative root-n consistent methods.

In addition to making these specific contributions concerning bootstrapped density estimators with empirically chosen bandwidths, we provide a detailed account of the way in which empirical bandwidth choice affects departure of the distribution of the density estimator from Normality. We also summarize numerical results that lend support to our conclusions, for both local and global plug-in rules. This work suggests an additional property: replacing $\hat{h}$ by $\hat{h}^*$ when constructing confidence intervals generally tends to increase coverage, regardless of whether this improves coverage accuracy or not.

Related work includes that of Taylor (1989) and Faraway and Jhun (1990) on bootstrap approaches to bandwidth selection, and Hall [(1992a,b, 1993)] on confidence regions based on nonparametric function estimators. The results in the present paper have of course direct analogues in the context of nonparametric regression, where similar difficulties arise with bootstrap estimators of bias.

Section 2 will summarize methodology, Section 3 will state our main theoretical results and discuss their implications, simulation results will be outlined in Section 4 and theoretical arguments for Section 3 will be given in Section 5.

2. Methodology.

2.1. *What is the correct location parameter?* We noted in Section 1 that standard bootstrap methods are unreliable for estimating biases of linear curve estimators. They generally estimate bias as zero, even when it is significant. Thus, standard bootstrap approaches to constructing confidence regions for curves should be interpreted as producing regions for the expected value of the estimator, rather than for the true curve.

However, if the bandwidth is chosen empirically then one should consider including the empirical bandwidth inside the expectation. In the present section we consider some of the difficulties raised by that approach and show how they may be overcome by truncation of the bandwidth estimator.

The following notation will be used throughout the paper. Given a random sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ from a univariate distribution with density f, we estimate $f(x)$ by

$$\hat{f}(x|h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

and put $\hat{f}_0 = \hat{f}(\cdot|h_0)$ and $\tilde{f} = \hat{f}(\cdot|\hat{h})$, where h_0 is a deterministic bandwidth and $\hat{h}$ is an empirical approximation to h_0. To appreciate the problems associated with moments of $\tilde{f}$, consider a bandwidth selector $\hat{h}$ that may be written as $\hat{h} = I_n \hat{h}_1 + (1 - I_n)\hat{h}_2$, where (1) the indicator variable I_n takes only the values 0 and 1, the latter with probability $p_n > 0$, and (2) conditional on I_n, $\hat{h}_1$ is uniformly distributed on the interval $(0, \delta_n)$, where $\delta_n \to 0$. Assume too that $K \geq 0$ and $K(0) > 0$. Then, $E\{\tilde{f}(X_i)\} = \infty$ for each i. Similarly, if $n^\alpha p_n$ and $n\delta_n$ are bounded away from 0, for some $\alpha > 0$, then the expected value of the supremum of $\tilde{f}(x)$ over any nondegenerate interval on which f is bounded away from 0 is unbounded.

This example illustrates the problems associated with taking expected values of estimators computed using stochastic bandwidths. These difficulties vanish, however, if it should be the case that $\hat{h}$ is bounded below by a constant multiple of n^{-A}, for some $A > 0$. Indeed, in such cases, $E(\tilde{f})$ may be calculated via a term-by-term Taylor series; see Lemma 5.1 in Section 5.

One could regard $f(x)$, rather than the expected value of an estimator, as the location parameter. A confidence region for the former can be constructed from one for the latter by making an explicit bias correction. Of course, this introduces a new layer of complications, above those discussed in the present paper. An alternative approach is to undersmooth when constructing $\hat{f}$, so that bias becomes less of an issue; see Hall (1992b) for discussion of these methods. However, there is a growing belief that the most appropriate approach to constructing confidence regions is to estimate $\hat{f}$ in a way that is optimal for pointwise accuracy, and construct a confidence interval or a pointwise confidence band for the expected value of this estimator. This view has been expressed particularly strongly in discussion on e-mail bulletin boards, where

it has been argued that such an approach has advantages of clarity, simplicity and easy interpretation.

2.2. *Local plug-in methods.* The asymptotically optimal bandwidth for estimating f at x, using $\hat{f}(x|h)$, is $h = h_0 = a_1\{f(x)/f''(x)^2\}^{1/5}n^{-1/5}$, where the constant a_1 depends only on K. See, for example, Wand and Jones [(1995), page 41]. To exposit properties of local plug-in methods we shall assume f has four continuous derivatives in a neighborhood of x, and $f(x)f''(x) \neq 0$; and consider estimating h_0 by

$$\hat{h} = a_1\{\hat{f}(x|h_1)/\hat{f}''(x|h_2)^2\}^{1/5}n^{-1/5}, \tag{2.1}$$

where $\rho h_1 \sim h_0$ for some constant $\rho > 0$ (the value of which would not be known to the experimenter), and $h_2 \asymp n^{-d}$, with $d < \frac{1}{5}$. Let $(\mathrm{A}_{\mathrm{lpi}})$ denote the intersection of these assumptions. Assuming $(\mathrm{A}_{\mathrm{lpi}})$ and taking $c = \min\{2d, \frac{1}{2}(1-5d)\}$, it may be proved that

$$P\bigl(|\hat{h}h_0^{-1} - 1| > n^{-c+\varepsilon}\bigr) = O\bigl(n^{-\lambda}\bigr) \text{ for all } \varepsilon, \lambda > 0. \tag{2.2}$$

The condition $d < \frac{1}{5}$ is needed to ensure that $\hat{f}''(x|h_2)$ is consistent for $f''(x)$. Taking $d = \frac{1}{9}$ minimizes the order of mean-squared error of $\hat{f}''(x|h_2)$ as an estimator of $f''(x)$, and so this is the order of the bandwidth that is typically recommended for calculating this part of $\hat{h}$. More generally, since h_2 is an order of magnitude larger than h_1, and K will be taken to be compactly supported, then the probability that $\hat{h} = 0$ may be shown to be strictly positive although exponentially small. Therefore, $E\{\tilde{f}(x)\}$ is not well defined if we take $\hat{h}$ as at (2.1). This difficulty may be overcome in a variety of ways; we shall threshold $\hat{h}$, obtaining $\hat{h}_{\mathrm{t}}$ defined by $\hat{h}_{\mathrm{t}} = \hat{h}$ if $\hat{h} > n^{-A}$ and $\hat{h}_{\mathrm{t}} = n^{-A}$ otherwise, where $A > 1/5$ but is otherwise arbitrary. For this definition of $\hat{h}_{\mathrm{t}}$ we put $\tilde{f}_{\mathrm{t}}(x) = \hat{f}(x|\hat{h}_{\mathrm{t}})$.

2.3. *Global plug-in methods.* These techniques are based on the fact that the optimal bandwidth for minimizing mean integrated squared error is asymptotic to $h_0 \equiv a_2J^{-1/5}n^{-1/5}$, where a_2 depends only on K, and $J = J(f) = \int(f'')^2$. See Scott [(1992), pages 130–131] and Wand and Jones [(1995), page 22]. There is a variety of root-n consistent estimators of J; we consider here

$$\widehat{J} = \frac{2}{n(n-1)h_3^5}\sum\sum_{1\le i\le j\le n} L^{(4)}\biggl(\frac{X_i - X_j}{h_3}\biggr), \tag{2.3}$$

where L is a new kernel and h_3 is a new bandwidth. Then,

$$\hat{h} \equiv a_2|\widehat{J}|^{-1/5}n^{-1/5} \tag{2.4}$$

is an empirical approximation to h_0. (The absolute value sign is used to remove any ambiguity, although from an asymptotic viewpoint it is unnecessary.)

We assume that L is a symmetric, compactly supported function with four bounded derivatives and satisfying

$$\int u^j L(u)\,du = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } j = 1, \ldots, 5, \end{cases}$$

f is compactly supported with eight bounded derivatives on the real line, and $h_3^{-1}n^{-1/9} + h_3 n^{1/12} = O(1)$. Call these conditions (A_{gpi}). The assumptions on f and L may be shown to imply that $\widehat{J} = J + O_p(h_3^6 + n^{-1/2} + n^{-1}h_3^{-9/2})$; see Lemma 3.1 of Hall and Marron (1987) for a closely related result. This property, and the assumptions on h_3 in (A_{gpi}), imply that $\widehat{J} = J + O_p(n^{-1/2})$. Therefore, ($A_{\text{gpi}}$) implies that $\hat{h}$ is "root-n consistent" for h_0, in the sense that $\hat{h}/h_0 = 1 + O_p(n^{-1/2})$.

Note that, since $|\widehat{J}| \le h_3^{-5} \sup |L^{(4)}|$, $\hat{h}$ is bounded below by a constant multiple of $n^{-14/45}$, with probability 1. This threshold makes the global plug-in bandwidth relatively resistant against unduly small choices, and ensures that $E(\tilde{f})$ is a good approximation to the "average" value of $\tilde{f}$, at least in an asymptotic sense.

It may be proved that, assuming (A_{gpi}), (2.2) holds with $c = \frac{1}{2}$.

2.4. *Estimating scale and constructing bootstrap approximations.* The variance of $\hat{f}(x|h)$ is $\sigma(x|h)^2 = (nh)^{-1}\gamma_2(x|f, h) - n^{-1}\gamma_1(x|f, h)^2$, where

$$\gamma_j(x|f, h) = \int K(u)^j f(x - hu)\,du.$$

A simple estimator of $\sigma(x|h)^2$ is

$$\hat{\sigma}(x|h)^2 = (nh)^{-1}\gamma_2(x|\bar{f}, h) - n^{-1}\gamma_1(x|\bar{f}, h)^2, \tag{2.5}$$

where $\bar{f}$ is an estimator of f. In the estimator $\hat{f}(\cdot|h)$ we would usually take $h = \hat{h}$, and so this choice of h would generally also be employed for $\hat{\sigma}(\cdot|h)$. Selection of $\bar{f}$ is not critical, but $\bar{f} = \tilde{f} = \hat{f}(\cdot|\hat{h})$ is an obvious choice, and except in the next paragraph, where we discuss more general options, we shall use this definition of $\bar{f}$ when defining $\hat{\sigma}(x|h)$.

To appreciate why selection of $\bar{f}$ is not so important, observe that for any reasonable choice of $\bar{f}$, in particular for $\bar{f} = \hat{f}(\cdot|\hat{h})$, we have $\gamma_j(x|\bar{f}, h) = \gamma_j(x|f, h) + O_p(n^{-2/5})$ for each j, and so $\hat{\sigma}(x|h)^2\sigma(x|h)^{-2} = 1 + O_p(n^{-2/5})$, this result continuing to hold if $h = \hat{h}$. In consequence, replacing $\sigma(x|\hat{h})$ by $\hat{\sigma}(x|\hat{h})$, when normalizing a density estimator, affects only terms of order $n^{-2/5}$ or higher in approximations to distributions. When $\hat{h}$ is the local plug-in bandwidth, this is of smaller order than changes that are caused through replacing h_0 by $\hat{h}$ in the density estimator itself; see Theorem 3.1. If $\hat{h}$ is the global plug-in bandwidth then estimation of f in the formula for $\sigma(x|h)$ affects first-order properties only in the standard way that is to be expected for Studentized estimators; see Theorem 3.2.

Direct bootstrap approximations to the distributions of $\tilde{f}$ or $\tilde{f}_t$ are generally obtained by replacing the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ by a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$, drawn by sampling from $\mathcal{X}$ with replacement. In particular, the bootstrap form of $\hat{f}$ is

$$\hat{f}^*(x|h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right).$$

The bootstrap version of $\hat{h}$ is defined analogously, replacing $\hat{f}$ by $\hat{f}^*$ in (2.1) (in the local plug-in case), to obtain $\hat{h}^*$, or (for the global plug-in rule) replacing each datum X_i by the respective X_i^* in (2.3), to obtain $\widehat{J}^*$, and then substituting $\widehat{J}^*$ for $\widehat{J}$ in (2.4), to obtain $\hat{h}^*$. We shall take $\bar{f}_1^* = \hat{f}^*(\cdot|\hat{h})$ and $\bar{f}_2^* = \hat{f}^*(\cdot|\hat{h}^*)$ to be our bootstrap versions of $\bar{f}$, and define the bootstrap version of $\hat{\sigma}(x|h)^2$ by analogy with (2.5): for $j = 1, 2$,

$$\hat{\sigma}_j^*(x|h)^2 = (nh)^{-1}\gamma_2(x|\bar{f}_j^*, h) - n^{-1}\gamma_1(x|\bar{f}_j^*, h)^2. \tag{2.6}$$

For brevity we shall give detailed results only for a particular percentile-t approach to bootstrapping, although properties of the percentile method will be outlined. Let $\mu(x) = E\{\hat{f}_0(x)\}$. We consider approximating the distribution of $T = \{\hat{f}(x|\hat{h}) - \mu(x)\}/\hat{\sigma}(x|\hat{h})$ by the conditional distribution of either

$$T_j^*(x) = \frac{\hat{f}^*(x|\hat{h}) - \hat{f}(x|\hat{h})}{\hat{\sigma}_j^*(x|\hat{h})} \quad \text{or} \quad U_j^*(x) = \frac{\hat{f}^*(x|\hat{h}^*) - \hat{f}(x|\hat{h})}{\hat{\sigma}_j^*(x|\hat{h}^*)}, \tag{2.7}$$

for $j = 1, 2$. If R^* denotes either $T_j^*(x)$ or $U_j^*(x)$, and $\hat{v}_\alpha$ is defined by $P(R^* \le \hat{v}_\alpha|\mathcal{X}) = \alpha$, for $0 < \alpha < 1$, then $(-\infty, \hat{f}(x|\hat{h}) - \hat{v}_{1-\alpha}\hat{\sigma}(x|\hat{h})]$ is a confidence interval for $\mu(x)$ with coverage probability approximately equal to α.

3. Main theoretical results.

3.1. *Theoretical approximations to distributions.* Our distribution approximations will be developed around one-term Edgeworth expansions of the distribution of $\hat{f}_0 - E(\hat{f}_0)$, discussed for example by Hall [(1991); 1992a, Section 4.4]. In the non-Studentized case such an expansion has the form

$$\begin{aligned} P\{\hat{f}_0(x) - \mu(x) \le \sigma_0(x)z\} &= \Phi(z) - (nh_0)^{-1/2}\tfrac{1}{6}\beta f(x)^{-1/2} \\ &\quad \times (z^2 - 1)\phi(z) + o\{(nh_0)^{-1/2}\}, \end{aligned} \tag{3.1}$$

uniformly in $-\infty < z < \infty$, where $\mu(x) = E\{\hat{f}_0(x)\}$ and $\sigma_0(x)^2 = \operatorname{var}\{\hat{f}_0(x)\}$ denote the mean and variance of the density estimator, Φ and ϕ are the standard Normal distribution and density functions, and $\beta = (\int K^3)/(\int K^2)^{3/2}$.

In formulating the Studentized form of (3.1), we define $\hat{\sigma}(x|h)$ by (2.5) with $\bar{f} = \hat{f}(\cdot|\hat{h})$. Then,

$$\begin{aligned} P\{\hat{f}_0(x) - \mu(x) \le \hat{\sigma}(x|h_0)z\} &= \Phi(z) + (nh_0)^{-1/2}\tfrac{1}{6}\beta\ f(x)^{-1/2} \\ &\quad \times (2z^2 + 1)\phi(z) + o\{(nh_0)^{-1/2}\}, \end{aligned} \tag{3.2}$$

again uniformly in $-\infty < z < \infty$.

The only additional regularity conditions needed for (3.1) and (3.2), apart from (A_{lpi}) or (A_{gpi}), are that (1) $f(x) \neq 0$, and (2) K is continuously differentiable and supported on a compact interval $\mathscr{K}$, and there are only a finite number of points in $\mathscr{K}$ where K vanishes. See, for example, Hall [(1992a), Section 5.5].

Let c equal $\min\{2d, \frac{1}{2}(1-5d)\}$ or $\frac{1}{2}$ in the cases of local plug-in or global plug-in, respectively. [These are the maximum values for which (2.2) holds.] We shall assume that K is a symmetric, compactly supported probability density, vanishing only at a finite number of points in its support interval and with ν bounded derivatives, where $\nu > (2/5c) + 1$ and $\nu = 2$ in the local and global plug-in cases, respectively. Call these assumptions (A).

Define $\kappa = \int K^2$, $\kappa_2 = \int u^2 K(u)\,du$,

$$\kappa(\rho) = \rho \int K(u)\{\rho u K'(\rho u) + K(\rho u)\}\,du,$$

$$\tau_1(x) = \frac{\kappa_2 f^{(4)}(x)}{5f''(x)}, \qquad \tau_2(x) = \frac{K''(0) f(x)}{5\kappa_2 f''(x)^2}.$$

Recall that $\tilde{f} = \hat{f}(\cdot|\hat{h})$, and that in the local plug-in case, ρ equals the limit of h_0/h_1 as $n \to \infty$. We use $\kappa(\rho)$ and $\tau_j(x)$ only in the local plug-in setting. Let $\operatorname{sgn} u$ denote the sign of a nonzero real number u.

THEOREM 3.1 (Local plug-in method). *Assume conditions* (A) *and* (A_{lpi}). *Then*

$$\begin{aligned} &P\{\tilde{f}(x) - \mu(x) \le \sigma_0(x) z\} \\ &\quad = \Phi(z) + h_2^2 \tau_1(x)\{\kappa(\rho)\kappa^{-1} z - \operatorname{sgn} f''(x)\}\phi(z) - h_0^{-2}(n h_2^3)^{-1}\tau_2(x) \\ &\qquad \times \{(z^2-1)\operatorname{sgn} f''(x) + 2z\}\phi(z) + o(h_2^2 + n^{-3/5} h_2^{-3}), \end{aligned} \tag{3.3}$$

uniformly in $-\infty < z < \infty$. *The effect of using the thresholded local plug-in bandwidth selector* $\hat{h}_t$ *and replacing* $\mu(x)$ *by* $E\{\tilde{f}_t(x)\}$ *in the probability on the left-hand side of* (3.3) *is to remove one of the terms in* $\operatorname{sgn} f''(x)$ *from the right-hand side*:

$$\begin{aligned} &P[\tilde{f}_t(x) - E\{\tilde{f}_t(x)\} \le \sigma_0(x) z] \\ &\quad = \Phi(z) + h_2^2 \tau_1(x)\kappa(\rho)\kappa^{-1} z\phi(z) - h_0^{-2}(n h_2^3)^{-1}\tau_2(x) \\ &\qquad \times \{(z^2-1)\operatorname{sgn} f''(x) + 2z\}\phi(z) + o(h_2^2 + n^{-3/5} h_2^{-3}), \end{aligned} \tag{3.4}$$

uniformly in z. *Results* (3.3) *and* (3.4) *continue to hold if* $\sigma_0(x) = \sigma(x|h_0)$ *is replaced by* $\hat{\sigma}(x|h_0)$ *within the probability statement on the left-hand side of each*. *Moreover*, (3.3) *and* (3.4) *hold if* $\sigma_0(x) = \sigma(x|h_0)$ *is replaced by either* $\sigma(x|\hat{h})$ *or* $\hat{\sigma}(x|\hat{h})$, *except that* (*for either of these choices of normalization*) *the term*

$$\{\tfrac{1}{2} h_2^2 \tau_1(x) + h_0^{-2}(n h_2^3)^{-1}\tau_2(x)\operatorname{sgn} f''(x)\} z\phi(z) \tag{3.5}$$

should be added to the right-hand side.

Result (3.4) does not change if, on the left-hand side, we replace $\sigma_0(x)$ by the standard deviation of $\tilde{f}_t(x)$.

REMARK 3.1 (Dominant errors in distribution expansions). Note that adding the term at (3.5) to the right-hand sides of (3.3) or (3.4) does not change the main character of either of those expansions: there are still two main terms in each, of sizes h_2^2 and $h_0^{-2}(nh_2^3)^{-1}$, respectively. Since h_2 is of strictly larger order than h_0, then the terms in h_2^2 are of strictly larger order than the terms in $(nh_0)^{-1/2}$ on the right-hand sides of (3.1) and (3.2). Therefore, the Edgeworth expansions (3.3) and (3.4), and their analogues with the quantity at (3.5) added, describe particularly low orders of departure from the asymptotic distribution. As we shall show in Section 3.2, the bootstrap has difficulty capturing such low-order departures.

REMARK 3.2 (Studentized versus non-Studentized cases). The fact that (3.3) and (3.4) are unchanged if $\sigma_0(x) = \sigma(x|h_0)$ is replaced by $\hat{\sigma}(x|h_0)$ reflects the fact that the impact of using the local plug-in bandwidth selector $\hat{h}$ is more significant than that of standardizing by the empirical standard deviation, provided the latter is computed using the optimal bandwidth h_0. This in turn is reflected by the fact that the difference between the non-Studentized and Studentized expansions (3.1) and (3.2) is of smaller order, $(nh_0)^{-1/2}$, than the dominant terms on the right-hand sides of (3.3) and (3.4).

REMARK 3.3 (Rate of convergence to asymptotic distribution). One of the implications of (3.3) and (3.4), and their versions with the term at (3.5) added, is that the distance of the distribution of $\tilde{f}$ (or $\tilde{f}_t$) from Normality is minimized by choosing h_2^2 and $n^{-3/5}h_2^{-3}$ to be of the same size; that is, by choosing $h_2 \asymp n^{-3/25}$. This is a different order from that (that is, $n^{-1/9}$; see Section 2.2) which minimizes the mean squared error of $\hat{f}''(x|h_2)$ as an estimator of $f''(x)$, although the difference is only slight.

REMARK 3.4 (Effects of oversmoothing or undersmoothing when choosing pilot bandwidths). The results discussed above assume $h_1 \asymp h_0$. If, however, h_1 is chosen to give a certain amount of oversmoothing and tends to zero more slowly than h_0, then, in effect, $\rho = 0$, and so $\kappa(\rho) = 0$. In consequence, the term in h_2^2 vanishes from the right-hand side of (3.4), although not from (3.3). Therefore, in the case of (3.4), faster rates of convergence to Normality may be achieved by oversmoothing when constructing the density estimator $\hat{f}(x|h_1)$ in the definition of $\hat{h}$ at (2.1).

When $0 < \rho < \infty$, the optimal rate of convergence to Normality in the supremum metric, obtained using $h_2 \asymp n^{-3/25}$, is $O(n^{-6/25})$. Again this choice of pilot bandwidth entails undersmoothing, relative to the "pointwise optimal" choice $h_2 \asymp n^{-1/9}$. The rate of convergence to Normality is only $O(n^{-2/9})$ in the latter case. These results should be compared with the rate $O(n^{-2/5})$ when the bandwidth is chosen non empirically; see (3.1) and (3.2).

Therefore, in principle, convergence rates can be improved by slightly oversmoothing when estimating f, or undersmoothing when estimating f'', in the formula for the pointwise-optimal bandwidth. However, practical implementation of such a method is far from straightforward. There are no existing techniques for choosing h_1 and h_2 against criteria of coverage accuracy, and the development of procedures is inhibited by the facts that (1) coverage accuracy needs to be optimized to second order, not first order, and (2) optimal values of h_1 and h_2 should depend on x. Difficulty (1) is reflected in the property that the ratio $3/25$ $(= 1/8.333\cdots)$, occurring in the convergence rate optimal formula $h_2 \asymp n^{-3/25}$, is very close to $1/9$, arising in the pointwise-optimal formula $h_2 \asymp n^{-1/9}$. By way of contrast, there are elementary well-known methods, such as the Normal scale approach (see Remark 3.6) and plug-in rules, for selecting h_1 and h_2 using pointwise-optimality criteria.

REMARK 3.5 [Explanation for terms on right-hand sides of (3.3)–(3.5)]. In Equations (3.3)–(3.5) the terms in h_2^2 come directly from the bias of $\hat{f}''(x|h_2)$, the latter quantity being used in $\hat{h}$ to estimate $f''(x_0)$ in the definition of h_0. The portion $(nh_2^3)^{-1}$ of the term in $h_0^{-2}(nh_2^3)^{-1}$ comes indirectly from the stochastic component of $\hat{f}''(x|h_2)$, arising through cross-products. The part h_0^{-2} comes from the normalization of the density estimator itself, appearing in the probabilities on the left-hand sides of (3.3)–(3.5); note that the normalization there is by the inverse of the standard deviation, and hence by the factor $(nh_0)^{1/2} \asymp h_0^{-2}$.

The global plug-in case is substantially different and more simple. There the empirical bandwidth selector $\hat{h}$ is so accurate, as an approximation to its "ideal" form h_0, that the first terms in the Edgeworth expansions at (3.1) and (3.2) are preserved.

THEOREM 3.2 (Global plug-in method). *Assume conditions* (A) *and* (A_{gpi}). *Then result* (3.1) *is unchanged if, in the probability on the left-hand side,* $\hat{f}_0(x) - \mu(x)$ *is replaced by* $\tilde{f}(x) - \mu(x)$ *or by* $\tilde{f}(x) - E\{\tilde{f}(x)\}$. *Neither is there any change if, within the probability statement on the left-hand side,* $\sigma_0(x)$ *is replaced by* $\sigma(x|\hat{h})$. *Moreover,* (3.2) *is unchanged if* $\hat{f}_0(x) - \mu(x)$ *on the left-hand side is replaced by* $\tilde{f}(x) - \mu(x)$ *or by* $\tilde{f}(x) - E\{\tilde{f}(x)\}$, *or if* $\hat{\sigma}(x|h_0)$ *on the left is replaced by* $\hat{\sigma}(x|\hat{h})$.

REMARK 3.6 (Effect of empirical choice of pilot bandwidth). Theorems 3.1 and 3.2 often remain true if the pilot bandwidths h_1, h_2 (used in the local plug-in case) and h_3 (used for global plug-in) are chosen empirically. For example, if we select h_1 using the "Normal scale" or "Normal reference" method [see, e.g., Silverman (1986), pages 45–47; Scott (1992), page 131; Wand and Jones (1995), page 60] and either select h_2 using the same principle, or, more simply, take $h_2 = h_1^{5/9}$; then Theorem 3.1 remains true provided the distribution with density f has finite fourth moment.

3.2. *Bootstrap approximations.* We confine attention almost solely to the local plug-in case, dealing with global plug-in methods only at the very end of the section. In the case of local plug-in bandwidth selectors, the main distinguishing feature of expansions in the bootstrap case is that they fail to capture bias contributions arising from the component $\hat{f}''(x|h_2)$ of $\hat{h}$. These are represented the terms in h_2^2 in (3.3)–(3.5).

Recall the definitions of $T_j^*(x)$ and $U_j^*(x)$ in Section 2.4, and note that $U_j^*(x)$ is the version of $T_j^*(x)$ in which $\hat{h}$ is replaced by $\hat{h}^*$ when computing bootstrap estimators. Therefore, conventional bootstrap arguments (see Section 1) would suggest that the bootstrap distribution of $U_j^*(x)$ should offer a better approximation, by an order of magnitude, than the bootstrap distribution of $T_j^*(x)$ to the true distribution of $T(x) = \{\tilde{f}(x) - \mu(x)\}/\hat{\sigma}(x|\hat{h})$. The distribution of the latter is treated in the later part of Theorem 3.1, where the term at (3.5) is to be added to the right-hand sides of (3.3) and (3.4), giving

$$\begin{aligned} P\{T(x) \le z\} &= \Phi(z) + h_2^2 \tau_1(x)\{\tfrac{1}{2}z + \kappa(\rho)\kappa^{-1}z - \operatorname{sgn} f''(x)\}\phi(z) \\ &\quad - h_0^{-2}(nh_2^3)^{-1}\tau_2(x)\{(z^2 - z - 1)\operatorname{sgn} f''(x) + 2z\}\phi(z) \\ &\quad + o(h_2^2 + n^{-3/5}h_2^{-3}). \end{aligned} \tag{3.6}$$

THEOREM 3.3 (Local plug-in method). *Assume conditions* (A) *and* (A_{lpi}). *Then*

$$\begin{aligned} P\{T_j^*(x) \le z|\mathcal{X}\} &= \Phi(z) - h_0^{-2}(nh_2^3)^{-1}\tau_2(x)\{(z^2 - 1)\operatorname{sgn} f''(x) + 2z\} \\ &\quad \times \phi(z) + o_p(h_2^2 + n^{-3/5}h_2^{-3}), \end{aligned} \tag{3.7}$$

$$\begin{aligned} P\{U_j^*(x) \le z|\mathcal{X}\} &= \Phi(z) - h_0^{-2}(nh_2^3)^{-1}\tau_2(x)\{(z^2 - z - 1)\operatorname{sgn} f''(x) + 2z\} \\ &\quad \times \phi(z) + o_p(h_2^2 + n^{-3/5}h_2^{-3}). \end{aligned} \tag{3.8}$$

uniformly in $-\infty < z < \infty$, *for both* $j = 1, 2$.

REMARK 3.7 [Explanation for terms on right-hand sides of (3.7) and (3.8)]. The terms in h_2^2 on the right-hand sides of expansions such as (3.3) and (3.6) come from the bias of $\hat{f}''(x|h_2)$, the latter appearing in the definition of $\hat{h}$. (See also Remark 3.5.) The nonappearance of terms in h_2^2 on the right-hand sides of (3.7) and (3.8), relative to (3.3) and (3.6), is due to the failure of the bootstrap distributions of either T_j^* or U_j^* to capture the bias of $\hat{f}''(x|h_2)$.

However, except for the term in h_2^2, the expansion at (3.7) is identical to that at (3.3). In particular, except for these terms, the bootstrap distribution of T_j^* captures the distributions of both $S_0(x) = \{\tilde{f}(x) - \mu(x)\}/\sigma_0(x)$ [see (3.3)] and $T_0(x) = \{\tilde{f}(x) - \mu(x)\}/\hat{\sigma}(x|h_0)$ [note that, by Theorem 3.1, (3.3) continues to hold if $\sigma_0(x)$ on its left-hand side is replaced by $\hat{\sigma}(x|h_0)$]. The fact that the distributions of both the non-Studentized statistic $S_0(x)$ and its Studentized

form $T_0(x)$ are captured to first order by the same bootstrap quantity, excepting the h_2^2 terms, reflects the property, noted in Remark 3.2, that at this level, Studentizing does not play a major role.

It might be thought that the bootstrap distribution of T_j^* should approximate the distribution of $Z = \{\hat f(x|h_0) - \mu(x)\}/\hat\sigma(x|h_0)$, as given at (3.2), and so the right-hand sides of (3.2) and (3.7) should be similar, except for terms in h_2^2. This is not the case, however, and (3.7) includes a term in $h_0^{-2}(nh_2^3)^{-1}$ that arises from stochastic error (as distinct from bias) of $\hat f''(x|h_2)$ in $\hat h$; note that Z does not involve $\hat h$.

REMARK 3.8 (Distributional approximation provided by U_j^*). Since the expansion (3.8) is identical to (3.6), except for terms in h_2^2, then apart from those terms, the bootstrap distribution of U_j^* captures the distributions of both $\{\tilde f(x) - \mu(x)\}/\sigma(x|\hat h)$ and $T(x) = \{\tilde f(x) - \mu(x)\}/\hat\sigma(x|\hat h)$. This again reflects the relatively unimportant role played by Studentizing.

REMARK 3.9 (Should $\mu(x)$ or $E\{\tilde f_{\rm t}(x)\}$ be considered as the target?). In Theorem 3.1 the only change brought about by considering $E\{\tilde f_{\rm t}(x)\}$, rather than $\mu(x)$, as the "mean" of the density estimator was to remove a portion of the term in h_2^2 from the Edgeworth expansion; compare (3.3) and (3.4). However, the h_2^2 term is not captured by the bootstrap distributions of either T_j^* or U_j^*, and so in the bootstrap setting there is not a clear argument for preferring $\mu(x)$ or $E\{\tilde f_{\rm t}(x)\}$ as the target of a confidence procedure.

By way of contrast, in the case of global plug-in bandwidth selectors, $\hat h$ is such an accurate approximation to h_0 that biases in the construction of $\hat h$ do not influence the distribution of the density estimator $\hat f(x|\hat h)$, to first order. This high level of accuracy is maintained by the bootstrap bandwidth selector $\hat h^*$, as our next result shows.

THEOREM 3.4 (Global plug-in method). *Assume conditions* (A) *and* ($\rm A_{gpi}$). *Then* (3.2) *is unchanged if the probability on the left-hand sides is changed to either* $P\{T_j^*(x) \le z|\mathcal{X}\}$ *or* $P\{U_j^*(x) \le z|\mathcal{X}\}$, *for either* $j = 1$ *or* 2, *provided the remainder term* $o\{(nh_0)^{-1/2}\}$ *on the right-hand side is replaced by* $o_p\{(nh_0)^{-1/2}\}$.

REMARK 3.10 (Percentile-method versions of Theorems 3.3 and 3.4). Like Theorem 3.3, Theorem 3.4 describes performance of a percentile-t method. Consider replacing $T_j^*(x)$ and $U_j^*(x)$, in the statement of Theorem 3.4, by their respective percentile forms,

$$V^* = \frac{\hat f^*(x|\hat h) - \hat f(x|\hat h)}{\hat\sigma(x|\hat h)} \quad \text{and} \quad W^* = \frac{\hat f^*(x|\hat h^*) - \hat f(x|\hat h)}{\hat\sigma(x|\hat h^*)},$$

where $\hat\sigma(x|h)$ is as defined in the first paragraph of Section 2.4. Then Theorem 3.4 continues to hold, provided we replace "(3.2)" by "(3.1)" in its formulation.

This is in effect a bootstrap version of a property first noted in Section 2.4: the global plug-in bandwidth selector $\hat{h}$ is so accurate that replacing $\hat{h}$ by its bootstrap form affects first-order properties only in the standard ways that are to be expected for Studentized or non-Studentized estimators.

By way of contrast, if in Theorem 3.3 we replace $T_j^*(x)$ and $U_j^*(x)$ by V^* and W^*, then there is no change to that result. This is another aspect of a property already noted in Remarks 3.7 and 3.8: the effects of local plug-in bandwidth selection dominate those of Studentizing.

4. Numerical properties.

4.1. *Parameters of the simulation study.* We used Monte Carlo simulation to evaluate small-sample properties of $\hat{f}^*(x|\hat{h})$ and $\hat{f}^*(x|\hat{h}^*)$. The underlying distributions were chosen to be four of the Normal mixture densities described by Marron and Wand (1992): (a) standard Normal, (b) skewed unimodal density [mixture of $N(0, 1)$, $N(1/2, (2/3)^2)$ and $N(13/12, (5/9)^2)$ in proportions 1:1:3], (c) bimodal density [equal mixture of $N(\pm 1, (2/3)^2)$], and (d) trimodal density [mixture of $N(-6/5, (3/5)^2)$, $N(6/5, (3/5)^2)$ and $N(0, (1/4)^2)$ in proportions 9:9:2]. Graphs of the densities are given by Marron and Wand (1992). Populations (a), (c) and (d) are symmetric about the origin, and that point was selected as a value of x for all four populations. We also chose $x = 0.75$ for population (a), $x = 1.00$ for populations (b) and (c) (in each case this point is close to a mode), and $x = 1.20$ for population (d) (again, close to a mode). Thus, simulations for each population were conducted using two values of x.

Four different bootstrap methods were considered in all instances, based on T_1^*, T_2^*, U_1^* and U_2^*, respectively; see (2.7) for definitions. Sample sizes were $n = 50$, 100 and 400; $B = 299$ bootstrap resamples were used to construct confidence intervals; and coverage probabilities were approximated by averaging over 3,000 replications. [We apportioned computational labor in this way since it is known that even taking B quite small has negligible impact on computational accuracy; see Hall (1986) and Hall and Titterington (1989).] Each confidence interval was one-sided and had nominal coverage 0.95, but in each setting (that is, for each population, each value of x, each sample size and each bootstrap method) we treated both left-handed $(-\infty, \hat{z}_{0.95})$ and right-handed $(\hat{z}_{0.05}, \infty)$ one-sided intervals. All density estimates were constructed using the standard Normal kernel, ϕ, the tails of which are so light that it is effectively compactly supported.

We computed the local plug-in bandwidth directly from (2.1), taking h_1 and h_2 there to be those bandwidths that would be (asymptotically) optimal for estimating f and f'', respectively, if the population were Normal with scale equal to the sample standard deviation. To implement the global plug-in rule we used the method described at (2.3) and (2.4), taking the kernel L to be of sixth order, $L(u) = \frac{1}{8}(u^4 - 10u^2 + 15)\phi(u)$. The bandwidth h_3 at (2.3) was

taken to be the one that would be optimal if the density were Normal with scale equal to the sample standard deviation.

4.2. *Results of the study.* The results reported below are all for the case where $\hat{f}(x|\hat{h})$ is treated as an estimator of $\mu(x) = E\{\hat{f}(x|h_0)\}$. The trends are the same if instead $E\{\hat{f}(x|\hat{h})\}$ is the focus of interest, except that all coverages tend to be slightly increased. As will be seen from the discussion below, this does not alter our conclusions.

There is a general tendency for coverage to increase if the empirical bandwidth $\hat{h}$ is replaced by $\hat{h}^*$ at the bootstrap step, regardless of whether a local or global plug-in method is used. In particular, for left-handed intervals, in only 1 out of 48 cases does the interval based on U_j^* not have at least the coverage of that based on T_j^*. For right-handed intervals the proportion is 7 out of 48. (The 48 cases arise as 4 populations $\times$ 3 sample sizes $\times$ 2 values of x $\times$ 2 plug-in methods.) Of the $1 + 7 = 8$ exceptions, all but two [these occurring in population (b)] arise in the case of the relatively complex populations (c) and (d). There, accurate estimation of f'', which is required for both local and global plug-in methods, is relatively difficult, and so it is perhaps not surprising that the trend is not followed completely.

Whether or not an increase in coverage is beneficial depends of course on the extent of the increase and on the base. In the case of left-handed confidence regions, where methods founded on T_j^* generally undercover, the more conservative performance evidenced by U_j^* often leads to enhanced coverage accuracy. For example, results for population (a) in the context of local plug-in smoothing, given in Table 1, show that there, U_1^* gives better results (in terms of closeness of true coverage to 0.95) than T_1^* in six out of six cases. The proportion is only three out of six for the global plug-in approach applied to population (a), however. Using the same criterion, and for populations (b), (c) and (d) together, the proportions in favor of U_1^* are 17 out of 18 for local plug-in smoothing, and 9 out of 18 for the global plug-in method.

On the other hand, methods based on T_j^* provide relatively good coverage accuracy in the case of right-handed confidence intervals, and so their counterparts employing U_j^* have a tendency to overcover to an unnecessarily large extent. Thus, for local plug-in smoothing and right-handed confidence regions, T_1^* outperforms U_1^* 15 out of 24 times, in terms of nearness to the nominal level. In the case of global plug-in methods, T_1^* outperforms U_1^* 19.5 out of 24 times. (The fraction comes from ties, which are counted 50% toward each tally. These figures are taken over all four populations.) Table 2, which shows the results for population (a) in the setting of local plug-in smoothing and right-handed confidence intervals, is fairly typical of properties in the case of unimodal populations. There, owing to relatively accurate coverage of intervals based on T_1^*, and consequent overcoverage in the case of intervals based on U_1^*, the former are more accurate in all six cases.

For both right- and left-handed intervals, results for U_2^* versus T_2^* tend to be more mixed and have less of an obvious trend. However, again there is

TABLE 1
*Left-handed confidence intervals in case of Normal population and local plug-in smoothing**

	$n=50$		$n=100$		$n=400$	
	$x=0.00$	$x=0.75$	$x=0.00$	$x=0.75$	$x=0.00$	$x=0.75$
$f(x)$	0.3989	0.3011	0.3989	0.3011	0.3989	0.3011
h_0	0.4267	0.6283	0.3714	0.5469	0.2815	0.4145
$E\hat{f}_0(x)$	0.3669	0.2761	0.3740	0.2819	0.3840	0.2899
p_{t1}	80.17	80.03	81.50	83.37	87.37	89.13
m_{t1}	0.4251	0.3207	0.4238	0.3170	0.4201	0.3137
s_{t1}	0.0864	0.0595	0.0636	0.0425	0.0332	0.0214
p_{t2}	89.87	86.47	92.90	90.63	97.60	94.70
m_{t2}	0.4969	0.3529	0.4673	0.3380	0.4338	0.3200
s_{t2}	0.1051	0.0743	0.0664	0.0476	0.0295	0.0207
p_{u1}	93.00	93.50	95.97	94.60	98.06	94.63
m_{u1}	0.4656	0.3501	0.4529	0.3351	0.4291	0.3180
s_{u1}	0.0819	0.0586	0.0571	0.0422	0.0298	0.0213
p_{u2}	97.66	95.20	98.67	96.30	99.76	96.83
m_{u2}	0.7102	0.4617	0.5894	0.4005	0.4584	0.3306
s_{u2}	0.2543	0.1420	0.1689	0.0927	0.0621	0.0310

*Values of $f(x)$, h_0 and $E\{\hat{f}_0(x)\}$ are given in the first set of three rows. Subsequent sets of three rows give Monte Carlo approximations to true coverage, $p(\times 100)$, average values of finite endpoints, m, and standard deviations of those endpoints, s, respectively, of the respective confidence regions. Subscripts tj and uj refer to local plug-in methods based on T_j^* and U_j^*, respectively.

no tendency for U_2^* to outperform T_2^*, or vice versa, in terms of sheer coverage accuracy.

We conclude that there is no evidence that methods based on U_j^* have systematically greater coverage accuracy than methods that use T_j^*. However, intervals based on U_j^* are generally more conservative, and so in cases where that is an advantage they would be preferable.

5. Derivations of theorems.

5.1. *Preliminary lemma.* Define $\Delta = h_0\hat{h}^{-1} - 1$. Let $c > 0$ be a constant such that (2.2) holds; let (A_1) denote the assumption that K is a symmetric, compactly supported probability density with $k_1 \geq 2$ bounded derivatives, f has two bounded derivatives in a neighborhood of x, and $\hat{h}$ satisfies (2.2); let (A_2) represent the assumption that for some $B_1, B_2 > 0$, $\hat{h} \geq B_1 n^{-B_2}$ with probability 1. Define $K_j(u) = u^j K^{(j)}(u)$ (for $j \geq 0$), $M_j(u) = K_j(u) + jK_{j-1}(u)$ (for $j \geq 1$) and $\hat{f}_j(x) = (nh_0)^{-1} \sum_i M_j\{(x - X_i)/h_0\}$ (for $j \geq 1$).

LEMMA 5.1. *For each $1 \leq k < k_1 - (2/5c)$ we may write*

$$\tilde{f}(x) = \hat{f}_0(x) + \Delta\hat{f}_1(x) + \cdots + \frac{1}{k!}\Delta^k\hat{f}_k(x) + R_1(x), \tag{5.1}$$

TABLE 2
*Right-handed confidence intervals in case of Normal population and local plug-in smoothing**

	$n=50$		$n=100$		$n=400$	
	$x=0.00$	$x=0.75$	$x=0.00$	$x=0.75$	$x=0.00$	$x=0.75$
$f(x)$	0.3989	0.3011	0.3989	0.3011	0.3989	0.3011
h_0	0.4267	0.6283	0.3714	0.5469	0.2815	0.4145
$E\hat{f}_0(x)$	0.3669	0.2761	0.3740	0.2819	0.3840	0.2899
p_{t1}	94.60	94.23	94.23	94.83	94.93	95.03
m_{t1}	0.2913	0.2233	0.3129	0.2388	0.3455	0.2623
s_{t1}	0.0466	0.0329	0.0377	0.0259	0.0235	0.0159
p_{t2}	89.70	89.87	90.67	91.67	93.13	93.73
m_{t2}	0.3039	0.2301	0.3220	0.2441	0.3499	0.2646
s_{t2}	0.0512	0.0385	0.0395	0.0277	0.0234	0.0159
p_{u1}	98.96	98.77	98.57	98.73	97.13	97.83
m_{u1}	0.2435	0.1853	0.2843	0.2156	0.3351	0.2551
s_{u1}	0.0631	0.0477	0.0490	0.0370	0.0262	0.0175
p_{u2}	97.26	96.60	96.53	97.40	96.10	96.80
m_{u2}	0.2792	0.2146	0.3042	0.2324	0.3415	0.2597
s_{u2}	0.0445	0.0324	0.0368	0.0256	0.0230	0.0155

*Rows have the same interpretation as before.

where, assuming (A) *and* (A_1), *we have that for some* $\varepsilon > 0$ *and all* $\lambda > 0$,

$$P\{|R_1(x)| > n^{-(2/5)-kc-\varepsilon}\} = O(n^{-\lambda}). \tag{5.2}$$

Assuming in addition (A_2),

$$E\{\tilde{f}(x)\} = E\left\{\hat{f}_0(x) + \Delta\hat{f}_1(x) + \cdots + \frac{1}{k!}\Delta^k\hat{f}_k(x)\right\} + O(n^{-(2/5)-kc-\varepsilon}) \tag{5.3}$$

for some $\varepsilon > 0$.

PROOF. The quantities $C_1, C_2, \ldots$ will denote positive constants. By Taylor expansion,

$$\begin{aligned}\tilde{f}(x) &= \frac{1+\Delta}{nh_0}\sum_{i=1}^{n} K\left\{\frac{x-X_i}{h_0}(1+\Delta)\right\} \\ &= \hat{f}_0(x) + \Delta\hat{f}_1(x) + \cdots + \frac{1}{(k_1-1)!}\Delta^{k_1-1}\hat{f}_{k_1-1}(x) + \Delta^{k_1}R_2(x),\end{aligned} \tag{5.4}$$

where, assuming $C_1h_0 \le \hat{h} \le C_2h_0$ [which, in view of (A_1), may be assumed true with probability $1 - O(n^{-\lambda})$ for all $\lambda > 0$] and K is supported on the interval $[-C_3, C_3]$,

$$|R_2(x)| \le C_4(nh_0)^{-1}\sum_{i=1}^{n} I(|x-X_i| \le C_2C_3h_0) = C_4R_3(x), \tag{5.5}$$

say. It may be proved using Bernstein's inequality that if C_5 is chosen sufficiently large then the probability that $|R_3(x)|$ exceeds C_5 equals $O(n^{-\lambda})$ for all $\lambda > 0$. Hence, in view of (A_1) we have for all $\varepsilon, \lambda > 0$,

$$P\{|\Delta^{k_1} R_2(x)| > n^{-k_1 c+\varepsilon}\} = O(n^{-\lambda}). \tag{5.6}$$

Repeated integration by parts shows that

$$\int u^r M_j(u)\,du = (-1)^j \frac{(r+j-1)!}{(r-1)!} \int u^r K(u)\,du$$

for $r \geq 0$ and $j \geq 1$, the right-hand side being interpreted as 0 if $r = 0$. In particular, $\int u^r M_j(u)\,du = 0$ for $r = 0, 1$, and so $E\{\hat{f}_j(x)\} = O(h^2)$ for $j = 1, \ldots, k_1 - 1$; call this result (R). (Here we have used the fact that f has two bounded derivatives.) Markov's inequality may be used to prove that the probability that $|\hat{f}_j(x) - E\{\hat{f}_j(x)\}|$ exceeds $n^{-(2/5)+\varepsilon}$ equals $O(n^{-\lambda})$ for all $\varepsilon, \lambda > 0$. From this property and (R) it follows that the probability that $|\hat{f}_j(x)|$ exceeds $n^{-(2/5)+\varepsilon}$ equals $O(n^{-\lambda})$ for all ε, λ. Results (5.1) and (5.2) now follow from (5.4) and (5.6).

Let ess($|\Delta|$) denote the essential supremum of Δ, and note that by (A_2), ess($|\Delta|$) $\leq B_3 n^{B_4}$ for some $B_3, B_4 > 0$. Therefore, if (A_2) holds then for any positive integer ℓ,

$$E(|\Delta|^\ell) \leq n^{-(c-\varepsilon)\ell} + (B_3 n^{B_4})^\ell P(|\Delta| > n^{-(c-\varepsilon)}).$$

By (A_1), the latter probability equals $O(n^{-\lambda})$ for all $\varepsilon, \lambda > 0$, and so $E(|\Delta|^\ell) = O(n^{-(c-\varepsilon)\ell})$ for all $\varepsilon > 0$. Therefore, for any random variable Z_n,

$$|E(\Delta^\ell Z_n)| \leq \{E(\Delta^{2\ell}) E(Z_n^2)\}^{1/2} = O\{n^{-(c-\varepsilon)\ell} (EZ_n^2)^{1/2}\}. \tag{5.7}$$

We shall apply this result to terms $\Delta^\ell \hat{f}_\ell(x)$ on the right-hand side of (5.1), and so we shall take $Z_n = \hat{f}_\ell(x)$. In this case we have, on the right-hand side of (5.7), $EZ_n^2 = O(h_0^4)$. Note too that by (5.5), if (A_2) holds then ess$\{|\Delta|^{k_1} R_2(x)\} \leq B_5 n^{B_6}$ for some $B_5, B_6 > 0$. Therefore, in view of (5.6), and using an argument similar to that leading to (5.7),

$$E\{\Delta^{k_1} R_2(x)\} = O(n^{-k_1 c+\varepsilon}) \tag{5.8}$$

for all $\varepsilon > 0$. Result (5.3) follows from (5.4), (5.7) and (5.8). This completes the proof of Lemma 5.1. □

In the local and global plug-in cases, Lemma 5.1 implies that

$$\tilde{f} = \hat{f}_0 + S_1 + R_4, \quad \tilde{f} = \hat{f}_0 + R_5, \tag{5.9}$$

respectively, where $S_1 = \Delta \hat{f}_1$, R_4 satisfies

$$P\{|R_4(x)| > n^{-(2/5)-c-\varepsilon}\} = O(n^{-\lambda}), \tag{5.10}$$

and R_5 satisfies

$$P\{|R_5(x)| > n^{-(4/5)-\varepsilon}\} = O(n^{-\lambda}), \tag{5.11}$$

both holding for some $\varepsilon > 0$ and all $\lambda > 0$. The lemma implies too that

$$E(\tilde{f}_{\mathrm{t}}) = E(\hat{f}_0) + E(S_1) + O\bigl(n^{-(2/5)-c-\varepsilon}\bigr), \qquad E(\tilde{f}) = E(\hat{f}_0) + O\bigl(n^{-(4/5)-\varepsilon}\bigr), \tag{5.12}$$

respectively, both holding for some $\varepsilon > 0$. Here and below, for simplicity we shall often suppress the argument x in quantities such as $\tilde{f}$, $\hat{f}_j$ and R_j.

5.2. *Proof of* (3.3). Put $\delta_{(0)}(x) = E\{\hat{f}(x|h_1)\} - f(x)$, $\delta_{(2)}(x) = E\{\hat{f}''(x|h_2)\} - f''(x)$, $\Delta_{(0)}(x) = \hat{f}(x|h_1) - E\{\hat{f}(x|h_1)\}$, $\Delta_{(2)}(x) = \hat{f}''(x|h_2) - E\{\hat{f}''(x|h_2)\}$. Put $c = \min\{2d, \frac{1}{2}(1-5d)\} < \frac{2}{5}$. It may be proved by Taylor expansion that

$$\Delta = \tfrac{1}{5}\bigl\{2(\delta_{(2)} + \Delta_{(2)})(f'')^{-1} - (\delta_{(0)} + \Delta_{(0)})f^{-1}\bigr\} + R_7, \tag{5.13}$$

where, in view of Markov's inequality,

$$P\bigl(|R_7| > n^{-2c+\varepsilon}\bigr) = O\bigl(n^{-\lambda}\bigr)$$

for all $\varepsilon, \lambda > 0$. From this result, (5.10) and the first identity at (5.9) we deduce that for all sufficiently large n, all $\varepsilon, \lambda > 0$, and $j = 1$,

$$P(\tilde{f} - \mu \le z) \begin{cases} \le \Psi_j\bigl(z + n^{-(2/5)-2c+\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr) \\ \ge \Psi_j\bigl(z - n^{-(2/5)-2c+\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr), \end{cases} \tag{5.14}$$

where $\mu = E(\hat{f}_0)$,

$$\Psi_1(z) = P\bigl[\hat{f}_0 - \mu - \tfrac{1}{5}\hat{f}_1\bigl\{2(\delta_{(2)} + \Delta_{(2)})(f'')^{-1} - (\delta_{(0)} + \Delta_{(0)})f^{-1}\bigr\} \le z\bigr],$$

and the "$O(n^{-\lambda})$" terms are of that order uniformly in $-\infty < z < \infty$.

Note that $n^{-2c} = O\{(h_2^2 + n^{-3/5}h_2^{-3})n^{-\varepsilon}\}$ for some $\varepsilon > 0$. Call this result (R′). It will allow us to show that the terms in $n^{-(2/5)-2c+\varepsilon}$ at (5.14) pass into the "small oh" remainders in (3.3) and (3.4).

Define $\delta_j = E(\hat{f}_j)$ and $\Delta_j = \hat{f}_j - \delta_j$, for $j \ge 0$, and put

$$\begin{aligned} T_1 &= \Delta_0 - \tfrac{1}{5}\Delta_1\bigl\{2\delta_{(2)}(f'')^{-1} - \delta_{(0)}f^{-1}\bigr\} - \tfrac{1}{5}\delta_1\bigl\{2\Delta_{(2)}(f'')^{-1} - \Delta_{(0)}f^{-1}\bigr\}, \\ T_2 &= \tfrac{1}{5}\Delta_1\bigl\{2\Delta_{(2)}(f'')^{-1} - \Delta_{(0)}f^{-1}\bigr\}, \qquad T_3 = T_1 - T_2, \\ \Psi_2(z) &= P\bigl[T_3 \le z + \tfrac{1}{5}\delta_1\bigl\{2\delta_{(2)}(f'')^{-1} - \delta_{(0)}f^{-1}\bigr\}\bigr]. \end{aligned} \tag{5.15}$$

Then, (5.14) for $j = 1$ is identical to that result for $j = 2$.

Since T_1 is the sum of independent random variables with zero mean, then an Edgeworth expansion of the distribution of T_1 is relatively easy to derive. We claim that an Edgeworth expansion of the standardized distribution of T_3 equals that of T_1 plus a term of size $n^{-3/5}h_2^{-3}$:

$$\begin{aligned} &P\bigl\{T_1 \le (\operatorname{var} T_1)^{1/2}z\bigr\} - P\bigl\{T_3 \le E(T_3) + (\operatorname{var} T_3)^{1/2}z\bigr\} \\ &\qquad = (nh_0)^{1/2}\bigl(nh_2^3\bigr)^{-1}\bar{\tau}_2\bigl(z^2 - 1\bigr)\phi(z) + o(\xi_0), \end{aligned} \tag{5.16}$$

uniformly in z, where (for each $\varepsilon \ge 0$) $\xi_\varepsilon = (h_2^2 + n^{-3/5}h_2^{-3})n^{-\varepsilon}$, and $\bar{\tau}_2 = K''(0)f^{1/2}/5\kappa^{1/2}f''$. [Note that $E(T_1) = 0$.]

Methods for deriving (5.16) are variants of those discussed by Hall [(1992a), Section 5.5]. Note in particular that if $|h_0 - h_1|/h_0$ is bounded away from 0 as $n \to \infty$ then the conditions imposed on K imply that

$$\sup_{|t_1|+\cdots+|t_4|>\varepsilon} \left| \int \exp\left\{ it_1 K\left(\frac{x-y}{h_0}\right) + it_2 K\left(\frac{x-y}{h_1}\right) + it_3 M_1\left(\frac{x-y}{h_0}\right) + it_4 K''\left(\frac{x-y}{h_2}\right)\right\} f(y)\,dy \right| \le 1 - C(\varepsilon)h_2, \tag{5.17}$$

where $C(\varepsilon)$ depends on f, K and x as well as ε and is strictly positive for each $\varepsilon > 0$. This result plays the role of Cramér's smoothing condition, and allows us to rigorously develop Edgeworth expansions of the distributions of T_1 and T_3. If, however, $|h_0 - h_1|/h_0$ converges to 0 then (5.17) can fail. This case can be treated separately, and then conventional arguments used to treat settings where $|h_0 - h_1|/h_0$ converges to 0 only along a subsequence. For the sake of brevity we shall show only that cumulant expansions are consistent with (5.16). We shall prove too that (5.16) continues to hold if $E(T_3)$ and $\operatorname{var}(T_3)$ on the left-hand side of (5.16) are replaced by $E(T_1)$ and $\operatorname{var}(T_1)$, respectively.

Observe that $|E(\Delta_1\Delta_{(0)})| = O\{(nh_0)^{-1}\}$ and

$$\begin{aligned} E(\Delta_1\Delta_{(2)}) &= \big(nh_2^3\big)^{-1} \int M_1(u)\, K''(h_0u/h_2) f(x - h_0u)\,du + O\big(n^{-1}\big) \\ &= O\big\{(nh_2^3)^{-1}(h_0/h_2)^2 + n^{-1}\big\} = O\big(n^{-7/5}h_2^{-5} + n^{-1}\big), \end{aligned} \tag{5.18}$$

where we have Taylor-expanded $K''(h_0u/h_2)$ about 0 and used the facts that $\int u^j M_1(u)\,du = 0$ for $j = 0, 1$ and that K has four bounded derivatives. It follows that

$$(nh_0)^{1/2} E(T_3 - T_1) = O\big\{\big(nh_2^5\big)^{-1} + n^{-2/5}\big\} = O(\xi_\varepsilon) \tag{5.19}$$

for some $\varepsilon > 0$. Also, $E\big\{\big(\Delta_1\Delta_{(2)}\big)^2\big\} = O\big\{(nh_0)^{-1}\big(nh_2^5\big)^{-1}\big\}$,

$$\big|E\big(\Delta_0\Delta_1\Delta_{(0)}\big)\big| + \big|E\big(\Delta_0\Delta_1\Delta_{(2)}\big)\big| = O\big\{(nh_0)^{-2} + (nh_0)^{-1}\big(nh_2^3\big)^{-1}\big\},$$

from which result (and related ones where Δ_0 on the left-hand side is replaced by Δ_1, $\Delta_{(0)}$ or $\delta_1\Delta_{(2)}$) it follows that

$$(nh_0)\,(\operatorname{var} T_3 - \operatorname{var} T_1) = O\big\{(nh_0)^{-1} + \big(nh_2^5\big)^{-1}\big\} = O(\xi_\varepsilon) \tag{5.20}$$

for some $\varepsilon > 0$.

Similar calculations show that

$$\sum_{j=0}^{3} \bigl|E\bigl(\Delta_0^j \Delta_1^{3-j} \Delta_{(0)}\bigr)\bigr| = O\bigl\{(nh_0)^{-2}\bigr\}, \qquad \bigl|E\bigl(\Delta_1 \Delta_{(2)}^3\bigr)\bigr| = O\bigl\{\bigl(nh_2^5\bigr)^{-2} h_0^2\bigr\},$$

$$\sum_{j=0}^{3} \bigl|E\bigl(\Delta_0^j \Delta_1^{3-j} \Delta_{(2)}\bigr)\bigr| = O\bigl\{(nh_0)^{-1}\bigl(nh_2^3\bigr)^{-1}\bigr\}, \qquad \bigl|E\bigl(\Delta_1^2 \Delta_{(2)}^3\bigr)\bigr| = O\bigl\{\bigl(n^3 h_0 h_2^8\bigr)^{-1}\bigr\},$$

$$\sum_{j_1=0}^{2} \sum_{j_2=0}^{2} \sum_{j_3=0}^{2} \sum_{j_4=0,2} \bigl|E\bigl(\Delta_{j_1} \Delta_{j_2} \Delta_{j_3} \Delta_{(j_4)}^2\bigr)\bigr| + \sum_{j=0,2} \bigl|E\bigl(\Delta_1^3 \Delta_{(j)}^3\bigr)\bigr|$$
$$= O\Bigl\{(nh_0)^{-3} + (nh_0)^{-2}\bigl(nh_2^5\bigr)^{-1} + (nh_0)^{-1}\bigl(nh_2^5\bigr)^{-1}\bigl(n^{-7/5} h_2^{-5} + n^{-1}\bigr)\Bigr\}.$$

[Bounding the $j = 2$ term in the last-written series involves an argument similar to that at (5.18).] Furthermore,

$$\begin{aligned}
&E\bigl(\Delta_0^2 \Delta_1 \Delta_{(2)}\bigr) \\
&\quad = (nh_0)^{-3} (nh_2^3)^{-1} n^2 \Biggl[\Biggl\{\int K\Bigl(\frac{x-y}{h_0}\Bigr)^2 f(y)\,dy\Biggr\}\Biggl\{\int M_1\Bigl(\frac{x-y}{h_0}\Bigr) K''\Bigl(\frac{x-y}{h_2}\Bigr) f(y)\,dy\Biggr\} \\
&\qquad + 2\Biggl\{\int K\Bigl(\frac{x-y}{h_0}\Bigr) M_1\Bigl(\frac{x-y}{h_0}\Bigr) f(y)\,dy\Biggr\}\Biggl\{\int K\Bigl(\frac{x-y}{h_0}\Bigr) K''\Bigl(\frac{x-y}{h_2}\Bigr) f(y)\,dy\Biggr\}\Biggr] \\
&\qquad + O\bigl\{(nh_0)^{-1}(nh_2^3)^{-1} n^{-\varepsilon}\bigr\} \\
&\quad = 2(nh_0)^{-1}(nh_2^3)^{-1}\Bigl(\int K M_1\Bigr) K''(0) f^2 + O\bigl\{(nh_0)^{-1}(nh_2^3)^{-1} n^{-\varepsilon}\bigr\},
\end{aligned}$$

for some $\varepsilon > 0$. [We use arguments similar to those at (5.18) to show that the first product within square brackets above makes a contribution that goes into the remainder.] Combining the results in this paragraph, and noting that $\int K M_1 = \frac{1}{2}\kappa$ and $\operatorname{var} T_3 = (nh_0)^{-1}\kappa f + O\{(nh_0)^{-1} n^{-\varepsilon}\}$, we deduce that

$$(\operatorname{var} T_3)^{-3/2}\bigl\{E\bigl(T_3^3\bigr) - E\bigl(T_1^3\bigr)\bigr\} = -6\,(nh_0)^{1/2}\,\bigl(nh_2^3\bigr)^{-1}\,\bar{\tau}_2 + O(\xi_\varepsilon)$$

for some $\varepsilon > 0$. From this result, (5.19) and (5.20) we deduce that

$$\begin{aligned}
&(\operatorname{var} T_3)^{-3/2}\bigl\{E(T_3 - ET_3)^3 - E(T_1 - ET_1)^3\bigr\} \\
&\qquad = -6\,(nh_0)^{1/2}\,\bigl(nh_2^3\bigr)^{-1}\,\bar{\tau}_2 + O(\xi_\varepsilon).
\end{aligned} \tag{5.21}$$

Differences between the cumulants of $(nh_0)^{1/2} T_1$ and $(nh_0)^{1/2} T_3$ of index 4 or more are of order ξ_ε for some $\varepsilon > 0$. Note too that if the differences of skewnesses (i.e., centered third moments) between two asymptotically Normal random variables (both standardized for location and scale) equals $\eta = \eta(n)$, if $\eta \to 0$, and if other cumulants differ only by $o(\eta)$, then formally, the difference between the two distributions equals $\frac{1}{6}\,\eta\,(1 - z^2)\,\phi(z) + o(\eta)$. See Hall [(1992a), pages 46–48]. Combining these properties with (5.19)–(5.21), we see that we have established formally, although not as yet rigorously, result (5.16).

That (5.16) continues to hold if $E(T_3)$ and $\operatorname{var}(T_3)$ on the left-hand side are replaced by $E(T_1)$ and $\operatorname{var}(T_1)$, respectively, follows from (5.19) and (5.20).

Next we show that, if we define $T_4 \equiv \Delta_0 - \frac{2}{5}\Delta_1\delta_{(2)}(f'')^{-1}$, then

$$P\{T_1 \le (\operatorname{var} T_1)^{1/2} z\} - P\{T_4 \le (\operatorname{var} T_4)^{1/2} z\} = o(\xi_0), \tag{5.22}$$

uniformly in z. [Note that $E(T_1) = E(T_4) = 0$.] It may be proved that for all $\varepsilon, \lambda > 0$,

$$P\bigl(|\Delta_1| > n^{-(2/5)+\varepsilon}\bigr) = O\bigl(n^{-\lambda}\bigr), \quad P\bigl(|\Delta_{(0)}| > n^{-(2/5)+\varepsilon}\bigr) = O\bigl(n^{-\lambda}\bigr). \tag{5.23}$$

If we define $T_5 \equiv \Delta_0 - \frac{2}{5}(\Delta_1\delta_{(2)} + \delta_1\Delta_{(2)})(f'')^{-1}$ then, in view of (5.23) and the fact that $|\delta_1| + |\delta_{(0)}| = O(n^{-2/5})$, we have

$$P(T_1 \le z) \begin{cases} \le P\bigl(T_5 \le z + n^{-(4/5)+\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr) \\ \ge P\bigl(T_5 \le z - n^{-(4/5)+\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr), \end{cases} \tag{5.24}$$

uniformly in z, for all $\varepsilon, \lambda > 0$. Arguments similar to those in the previous paragraph show that the cumulants of $(\operatorname{var} T_4)^{-1/2} T_4$ and $(\operatorname{var} T_5)^{-1/2} T_5$ differ only in terms equal to $o(\xi_0)$. Result (5.22) follows from this property and (5.24), noting the arguments outlined in the paragraph subsequent to (5.16).

The cumulants of $(\operatorname{var} T_4)^{-1/2} T_4$ and $(\operatorname{var} \Delta_0)^{-1/2}\Delta_0$ also differ only in terms equal to $o(\xi_0)$. Therefore, (5.22) implies

$$P\{T_1 \le (\operatorname{var} T_1)^{1/2} z\} - P\{\Delta_0 \le (\operatorname{var} \Delta_0)^{1/2} z\} = o(\xi_0), \tag{5.25}$$

uniformly in z. From (5.16) and (5.25) (the version of the former having the mean and variance of T_3 replaced by those quantities for T_1), we deduce that, since $E(T_1) = 0$,

$$\begin{aligned} P\{T_3 \le (\operatorname{var} T_1)^{1/2} z\} &= P\{\Delta_0 \le (\operatorname{var} \Delta_0)^{1/2} z\} \\ &\quad - (nh_0)^{1/2} \bigl(nh_2^3\bigr)^{-1} \bar{\tau}_2 (z^2 - 1)\, \phi(z) + o(\xi_0). \end{aligned} \tag{5.26}$$

Note too that $(nh_0)(\operatorname{var} T_1 - \operatorname{var} T_5) = o(\xi_0)$, and

$$\begin{aligned} (nh_0)(\operatorname{var} T_5 - \operatorname{var} T_4) &= -\frac{4\, nh_0\, \delta_1}{5\, f''}\, E(\Delta_0\Delta_{(2)}) + o(\xi_0) \\ &= \frac{4}{5}\kappa_2\, K''(0)\, f\, nh_0^3 \bigl(nh_2^3\bigr)^{-1} + o(\xi_0), \end{aligned}$$

the last identity following from the results

$$\delta_1 = -f''\, h_0^2\, \kappa_2 + o\bigl(h_0^2\bigr), \qquad E\bigl(\Delta_0\Delta_{(2)}\bigr) = \bigl(nh_2^3\bigr)^{-1} K''(0)\, f + o\bigl\{\bigl(nh_2^3\bigr)^{-1}\bigr\}.$$

Also,

$$\begin{aligned} (nh_0)(\operatorname{var} T_4 - \operatorname{var} \Delta_0) &= -\frac{4\, nh_0\, \delta_{(2)}}{5\, f''}\, E(\Delta_0\Delta_1) + o(\xi_0) \\ &= -\frac{2}{5}\, \frac{\kappa(\rho)\, \kappa_2\, f\, f^{(4)}}{f''}\, h_2^2 + o(\xi_0), \end{aligned}$$

the last identity following from the results $\delta_{(2)} = \frac{1}{2} h_2^2 \kappa_2 f^{(4)} + o(h_2^2)$ and

$$nh_0 E(\Delta_0 \Delta_1) f^{-1} = \int K(u) M_1(uh_0/h_1)\, du + o(1) = \kappa(\rho) + o(1).$$

Furthermore, $\operatorname{var} \Delta_0 \sim (nh_0)^{-1} \kappa f$. Combining the results in this paragraph to this point we see that

$$\frac{\operatorname{var} T_1 - \operatorname{var} \Delta_0}{\operatorname{var} \Delta_0} = 2a + o(\xi_0), \tag{5.27}$$

where

$$a = \frac{1}{5} \left\{ \frac{2 \kappa_2 K''(0)}{\kappa} nh_0^3 (nh_2^3)^{-1} - \frac{\kappa(\rho) \kappa_2 f^{(4)}}{\kappa f''} h_2^2 \right\}.$$

Combining (5.26) and (5.27) we deduce that

$$\begin{aligned} P\{T_3 \le (\operatorname{var} \Delta_0)^{1/2} z\} &= P\{\Delta_0 \le (\operatorname{var} \Delta_0)^{1/2} z\} - (nh_0)^{1/2} (nh_2^3)^{-1} \\ &\quad \times \bar\tau_2 (z^2 - 1) \phi(z) - az \phi(z) + o(\xi_0). \end{aligned} \tag{5.28}$$

Recall that $\delta_1 = -f'' h_0^2 \kappa_2 + o(h_0^2)$ and $\delta_{(2)} = \frac{1}{2} h_2^2 \kappa_2 f^{(4)} + o(h_2^2)$. Using these results, (5.14) with $j = 2$, result (R′) [defined below (5.14)], and the properties $(nh_0^5)^{1/2} \to (\kappa f)^{1/2}/(\kappa_2 |f''|)$ and $\operatorname{var} \Delta_0 \sim (nh_0)^{-1} \kappa f$, where $\kappa = \int K^2$, we deduce that

$$\begin{aligned} P\{\tilde f - \mu \le (\operatorname{var} \Delta_0)^{1/2} z\} &= P\big(T_3 \le (\operatorname{var} \Delta_0)^{1/2} \big[z + \tfrac{1}{5} \delta_1 (\operatorname{var} \Delta_0)^{-1/2} \\ &\qquad \times \{ 2 \delta_{(2)} (f'')^{-1} - \delta_{(0)} f^{-1} \} \big] \big) + o(\xi_0) \\ &= P\big[T_3 \le (\operatorname{var} \Delta_0)^{1/2} \{ z - \tfrac{2}{5} \kappa_2 h_0^2 (\operatorname{var} \Delta_0)^{-1/2} \delta_{(2)} \} \big] + o(\xi_0) \\ &= P\{ T_3 \le (\operatorname{var} \Delta_0)^{1/2} (z - \tfrac{1}{5} \kappa_2 |f''|^{-1} f^{(4)} h_2^2) \} + o(\xi_0) \\ &= \Phi(z) - (nh_0)^{1/2} (nh_2^3)^{-1} \bar\tau_2 (z^2 - 1) \phi(z) \\ &\quad - (\tfrac{1}{5} \kappa_2 |f''|^{-1} f^{(4)} h_2^2 + az) \phi(z) + o(\xi_0), \end{aligned} \tag{5.29}$$

where we have explicitly used (5.28) to obtain the last line and implicitly used (5.28) in earlier steps to obtain a modulus of continuity for the probability approximations. Since $(nh_0)^{1/2} \bar\tau_2 = h_0^{-2} s \tau_2$ and $\frac{1}{5} \kappa_2 K''(0) \kappa^{-1} nh_0^3 = h_0^{-2} \tau_2$, where $s = \operatorname{sgn} f''$, then (5.29) is equivalent to (3.3).

5.3. *Proof of* (3.4). If, in the construction of $\tilde f$, we replace $\hat h$ by its thresholded form $\hat h_{\mathrm t}$, thereby obtaining the estimator $\tilde f_{\mathrm t}$ (see Section 2.2), then the first part of (5.12) holds. Moreover, we may simplify the term $E(S_1) = E(\Delta \hat f_1)$ in that formula by replacing Δ by the nonremainder portion of the right-hand side of (5.13). This allows us to show that for some $\varepsilon > 0$,

$$\begin{aligned} E(S_1) &= \tfrac{1}{5} E\Big[\{ 2 (\delta_{(2)} + \Delta_{(2)}) (f'')^{-1} - (\delta_{(0)} + \Delta_{(0)}) f^{-1} \} \hat f_1 \Big] + O(n^{-(2/5)-2c}) \\ &= \tfrac{1}{5} \delta_1 \{ 2 \delta_{(2)} (f'')^{-1} - \delta_{(0)} f^{-1} \} + O(n^{-(2/5)-\varepsilon} h_0^2) \end{aligned}$$

for some $\varepsilon > 0$. The nonremainder term in the last line is identical to the added "correction" on the right-hand side in the probability at (5.15). Therefore, if we replace $\tilde{f}$ and μ by $\tilde{f}_{\mathrm{t}}$ and $E(\tilde{f}_{\mathrm{t}})$, respectively, in the sequence of steps at (5.29), then they simplify to

$$\begin{aligned} P\{\tilde{f}_{\mathrm{t}} - E(\tilde{f}_{\mathrm{t}}) \le (\mathrm{var}\ \Delta_0)^{1/2} z\} &= P\{T_3 \le (\mathrm{var}\ \Delta_0)^{1/2} z\} + o(\xi_0) \\ &= \Phi(z) - (nh_0)^{1/2} (nh_2^3)^{-1} \bar{\tau}_2 (z^2 - 1) \\ &\quad \times \phi(z) - az\phi(z) + o(\xi_0), \end{aligned} \tag{5.30}$$

the last identity following directly from (5.28). Formula (5.30) is equivalent to (3.4). Similar arguments may be used to prove that (3.3) and (3.4) are unchanged if $\sigma_0(x) = \sigma(x|h_0)$ is replaced by $\hat{\sigma}(x|h_0)$ on the respective left-hand sides.

5.4. *Proof of version of* (3.3) *when* $\sigma(x|\hat{h})$ *replaces* $\sigma_0(x)$. Write σ_0 and $\tilde{\sigma}$ for $\sigma_0(x)$ and $\sigma(x|\hat{h})$, respectively. It may be proved that $(\tilde{\sigma}^2 - \sigma_0^2)/\sigma_0^2 = \Delta + R_8$, where R_8 satisfies

$$P\{|R_j| > n^{-2c_1+\varepsilon}\} = O(n^{-\lambda}) \tag{5.31}$$

for all $\varepsilon, \lambda > 0$. From this result and (5.13) we may show that $\tilde{\sigma}/\sigma_0 = 1 + \frac{1}{5}(\delta_{(2)} + \Delta_{(2)})(f'')^{-1} + R_9$, where R_9 satisfies (5.31). Therefore, replacing var Δ_0 by $\tilde{\sigma}^2$ on the far left-hand side of (5.29), and following the string of identities as before, we deduce that

$$\begin{aligned} P(\tilde{f} - \mu \le \tilde{\sigma} z) &= P\Big[T_3 \le (\mathrm{var}\,\Delta_0)^{1/2} \Big\{1 + \tfrac{1}{5}(\delta_{(2)} + \Delta_{(2)})(f'')^{-1}\Big\} \\ &\qquad \times \Big\{z - \tfrac{1}{5}\kappa_2 |f''|^{-1} f^{(4)} h_2^2\Big\}\Big] + o(\xi_0) \\ &= P\Big[T_3 \le (\mathrm{var}\ \Delta_0)^{1/2} \Big\{1 + \tfrac{1}{5}\Delta_{(2)}(f'')^{-1}\Big\} \\ &\qquad \times \Big\{z - \tfrac{1}{5}\kappa_2 |f''|^{-1} f^{(4)} h_2^2\Big\}\Big] + \tfrac{1}{2} h_2^2 \tau_1 z \phi(z) + o(\xi_0) \\ &= P\Big[T_3 \le (\mathrm{var}\,\Delta_0)^{1/2} \Big\{z - \tfrac{1}{5}\kappa_2 |f''|^{-1} f^{(4)} h_2^2\Big\}\Big] \\ &\qquad + \tfrac{1}{5}\sigma_0^{-1}(nh_2^3)^{-1} K''(0) \frac{f}{f''} z\phi(z) + \tfrac{1}{2} h_2^2 \tau_1 z \phi(z) + o(\xi_0) \\ &= P\Big[T_3 \le (\mathrm{var}\,\Delta_0)^{1/2} \Big\{z - \tfrac{1}{5}\kappa_2 |f''|^{-1} f^{(4)} h_2^2\Big\}\Big] \\ &\qquad + \Big\{h_0^{-2} (nh_2^3)^{-1} \tau_2 \,\mathrm{sgn}(f'') + \tfrac{1}{2} h_2^2 \tau_1\Big\} z\phi(z) + o(\xi_0). \end{aligned}$$

Thus, the expansion is identical to that at (3.3), up to terms of order $o(\xi_0)$, except that the term at (3.5) should be added to the right-hand side.

The case where $\hat{\sigma}(x|\hat{h})$ replaces $\sigma_0(x)$ is similar; note remarks in Section 2.4.

5.5. *Proof of Theorem* 3.2 (*global plug-in method*). Let X have the distribution of a generic X_i, and define $A(u) = E[L^{(4)}\{(u-X)/h_3\}]$, $\alpha = E\{A(X)\}$, $B_1(u, v) = L^{(4)}\{(u-v)/h_3\} - A(u) - A(v) + \alpha$ and $B_2(u) = A(u) - \alpha$. A standard U-statistic decomposition of $\widehat{J}$ is $\widehat{J} = \widehat{J}_1 + 2\widehat{J}_2 + \alpha$, where

$$\widehat{J}_1 = \frac{2}{n(n-1)h_3^5} \sum\sum_{1 \le i < j \le n} B_1(X_i, X_j), \qquad \widehat{J}_2 = \bigl(nh_3^5\bigr)^{-1} \sum_{i=1}^{n} B_2(X_i).$$

The 2ℓth moments of $\widehat{J}_1$ and $\widehat{J}_2$ may be shown to equal $O(n^{-\ell})$ for each integer $\ell \ge 1$. Also, $\alpha - J = O(n^{-1/2})$. Therefore, using Markov's inequality it may be proved that for each $\varepsilon, \lambda > 0$, $\Delta = (\widehat{J}/J)^{1/5} - 1$ satisfies

$$P\bigl(|\Delta| > n^{-(1/2)+\varepsilon}\bigr) = O\bigl(n^{-\lambda}\bigr).$$

From this result, (5.11) and the second identity in (5.9) we deduce that for all sufficiently large n, some $\varepsilon > 0$ and all $\lambda > 0$,

$$P(\tilde{f} - \mu \le z) \begin{cases} \le P\bigl(\hat{f}_0 - \mu \le z + n^{-(4/5)-\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr), \\ \ge P\bigl(\hat{f}_0 - \mu \le z - n^{-(4/5)-\varepsilon}\bigr) + O\bigl(n^{-\lambda}\bigr), \end{cases} \tag{5.32}$$

where the "$O(n^{-\lambda})$" terms are of that order uniformly in z.

Recall that $\sigma_0^2 = \operatorname{var} \hat{f}_0$. Edgeworth expansion of the distribution of $(\hat{f}_0 - \mu)/\sigma_0$, up to and including a term of size $n^{-2/5}$ and with remainder of smaller order, is given at (3.1). In view of (5.32) the same expansion applies to the distribution of $Q \equiv (\tilde{f} - \mu)/\sigma_0$. Also, the second result at (5.12) implies that we may replace μ by $E(\tilde{f})$ in the definition of Q without affecting the expansion up to terms of smaller order than $n^{-2/5}$. The fact that we may replace $\hat{\sigma}(x|h_0)$ by $\hat{\sigma}(x|\hat{h})$, without affecting (3.2), follows by Taylor expansion and the delta method, from the properties $\hat{\sigma}(x|\hat{h})/\sigma(x|h_0) = 1 + O_p(|h_0^{-1}\hat{h} - 1|)$ and $\hat{h}/h_0 = 1 + O_p(n^{-1/2})$. Theorem 3.2 is a consequence of these results.

5.6. *Notes on proofs of Theorems* 3.3 *and* 3.4. Bootstrap versions of Theorems 3.2 and 3.3 may be derived along the same lines as before, there being no difference (at the level of first-order terms) in those components of Edgeworth expansions that derive from the differences between bootstrap quantities and their conditional expected values. Note, for example, that conditional on $\mathcal{X}$, $\hat{f}^*(x|\hat{h}) - E\{\hat{f}^*(x|\hat{h})|\mathcal{X}\}$, $\hat{f}^*(x|h_1) - E\{\hat{f}^*(x|h_1)|\mathcal{X}\}$ and $(\hat{f}^*)''(x|h_2) - E\{(\hat{f}^*)''(x|h_2)|\mathcal{X}\}$ are jointly asymptotically Normal with zero mean and the same respective asymptotic covariances as $\hat{f}(x|h_0) - E\{\hat{f}(x|h_0)\}$, $\hat{f}(x|h_1) - E\{\hat{f}(x|h_1)\}$ and $\hat{f}''(x|h_2) - E\{\hat{f}''(x|h_2)\}$.

However, $E\{\hat{f}^*(x|\hat{h})|\mathcal{X}\}$, $E\{\hat{f}^*(x|h_1)|\mathcal{X}\}$ and $E\{(\hat{f}^*)''(x|h_2)|\mathcal{X}\}$ are respectively identical (with probability 1) to $\hat{f}(x|\hat{h})$, $\hat{f}(x|h_1)$ and $\hat{f}''(x|h_2)$, respectively. In particular, the biases are 0, and so terms in Edgeworth expansions that take the value 0 if biases vanish, are no longer present. For example,

working through the proof of (3.3) we find that

$$\begin{aligned} &P\{\hat f^*(x|\hat h^*) - \hat f(x|\hat h) \le \hat\sigma(x|\hat h)z\} \\ &\quad = -\hat h^{-2}\,(nh_2^3)^{-1}\hat\tau_2(x)\,\{(z^2-1)\operatorname{sgn}\hat f''(x|h_2) + 2z\}\phi(z) \\ &\qquad + o_p(h_2^2 + n^{-3/5}h_2^{-3}), \end{aligned} \tag{5.33}$$

where

$$\hat\tau_2(x) = \frac{K''(0)\,\hat f(x|\hat h)}{5\,\kappa_2\,\hat f''(x|h_2)^2}.$$

Formula (5.33) is the bootstrap version of (3.3), except that terms in h_2^2, which derive from bias terms, are not present.

Noting that $\hat f(x|\hat h)$, $\hat f''(x|h_2)$ and $\hat h/h_0$ converge in probability to $f(x)$, $f''(x)$ and 1, respectively, and in particular that $\hat\tau(x) \to \tau(x)$, we deduce from (5.33) that (3.3) continues to hold if on the left-hand side there we place the probability that appears on the left-hand side of (5.33), and if we interpret the remainder in (3.3) as being of the stated order "in probability." This leads to (3.7). The same argument, with the same interpretation of the remainder, shows that (3.6) continues to hold if on the left-hand side we place the probability that appears on the left-hand side of (3.8). This result is equivalent to (3.8).

Acknowledgments. This paper was written while the second author was supported by an Australian Research Council grant in the Centre for Mathematics and its Applications at the Australian National University. Both authors are grateful to two referees and an Associate Editor for their helpful comments.

REFERENCES

DAVISON, A. C. and HINKLEY, D. J. (1997). *Bootstrap Methods and Their Applications*. Cambridge Univ. Press

EFRON, B. and TIBSHIRANI, B. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

FARAWAY, J. J. and JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* ***85*** 1119–1122.

HALL, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Ann. Statist*. **14** 1453–1462.

HALL, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Math. Operat. Statistik Ser. Statist*. **22** 215–232.

HALL, P. (1992a). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

HALL, P. (1992b). Effect of bias estimation on coverage accuracy on bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.

HALL, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **55** 291–304.

HALL, P. and MARRON, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett*. **6** 109–115. [Correction (1988) *Statist. Probab. Lett.* **7** 87.]

HALL, P. and TITTERINGTON, D. M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *J. Roy. Statist. Soc. Ser. B* **51** 459–467.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist*. **20** 712–736.

SCOTT, D. J. (1992). *Multivariate Density Estimation—Theory, Practice, and Visualization*. Wiley, New York.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

TAYLOR, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76** 705–712.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

CENTRE FOR MATHEMATICS AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA A.C.T. 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au

DEPARTMENT OF STATISTICS
HANKUK UNIVERSITY OF FOREIGN STUDIES
YONGIN, KYOUNGGI 449–791
KOREA