Effects of Smoothing on Distribution Approximations

Peter Hall
Australian National University

and

Xiao-Hua Zhou
Indiana University School of Medicine

Abstract

We show that a number of apparently disparate problems, involving distribution approximations in the presence of discontinuities, are actually closely related. One class of such problems involves developing bootstrap approximations to the distribution of a sample mean when the sample includes both ordinal and continuous data. Another class involves smoothing a lattice distribution so as to overcome rounding errors in the normal approximation. A third includes kernel methods for smoothing distribution estimates when constructing confidence bands. Each problem in these classes may be modelled in terms of sampling from a mixture of a continuous and a lattice distribution. We quantify the proportion of the continuous component that is sufficient to "smooth away" difficulties caused by the lattice part. The proportion is surprisingly small — it is only a little larger than $n^{-1} \log n$, where n denotes sample size. Therefore, very few continuous variables are required in order to render a continuity correction unnecessary. The implications of this result in the problem of sampling both ordinal and continuous data are discussed, and numerical aspects are described through a simulation study. The result is also used to characterise bandwidths that are appropriate for smoothing distribution estimators in the confidence band problem. In this setting an empirical method for bandwidth choice is suggested, and a particularly simple derivation of Edgeworth expansions is given.

Keywords: Bandwidth, bootstrap, confidence band, confidence interval, continuity correction, coverage error, Edgeworth expansion, kernel methods, mixture distribution.

1 Introduction

1.1 Smoothing in distribution approximations

Rabi Bhattacharya has made very substantial contributions to our understanding of normal approximations in statistics and probability. None has been less important and influential than his exploration and application of smoothing as it is related to distribution approximations. For example, his development of ways of smoothing multivariate characteristic functions lies at the heart of his pathbreaking work on Berry-Esseen bounds and other measures of rates of

convergence in the multivariate central limit theorem (e.g. Bhattacharya 1967, 1968, 1970; Bhattacharya and Rao, 1976). His introduction of what has become known as the "smooth function model" (Bhattacharya and Ghosh, 1978), for describing properties of Edgeworth expansions of statistics that can expressed as smooth functions of means, has allowed wide-ranging asymptotic studies of statistical methods such as those based on the bootstrap. The present paper is a very small contribution, but nevertheless in a related vein — a small token of our appreciation of the considerable contribution that Rabi has made to distribution approximations in mathematical statistics.

A key assumption in many distribution approximations in statistics is that the distribution being approximated is continuous. Without this property, not only are approximation errors likely to be large, but special features that the approximations are often assumed to enjoy can be violated. These include the property that the coverage error of a two-sided confidence interval is an order of magnitude less than that for its one-sided counterpart. In a range of practical problems the assumption of smoothness can be invalid, however. In such cases there may sometimes be enough "residual" smoothing present in other aspects of the problem for it to be unnecessary to smooth in an artificial way. Nevertheless, even in these circumstances it is important to know how much residual smoothing is required, so that the adequacy of the residual smoothing can be assessed. In other problems there is simply not enough smoothing to overcome the most serious discretisation errors; there, artificial smoothing, for example using kernel methods, can be efficacious.

In the present paper, motivated by particular problems of both these types, we derive a general theoretical benchmark for the level of smoothing that is adequate in each case. In the first class of problem, encountered in several practical settings, we suggest an empirical method for assessing whether the benchmark has been attained. In the second class, related to smoothed distribution estimation, we introduce an empirical technique for determining how much smoothing should be provided. Both types of problem have a common basis, in that they represent mixture-type sampling schemes where a portion of the data are drawn from a smooth distribution and the remainder from a lattice distribution.

It is shown that the sampling fraction of the smooth component can be surprisingly small before difficulties arise through the roughness of the other component. The threshold is approximately $n^{-1} \log n$, where n denotes sample size. In the case of the second problem this result may be interpreted as a prescription for bandwidth choice, which can be implemented in practice using a smoothed bootstrap method. For the first problem the result may be interpreted as defining a safeguard: only when the smooth component is present in a particularly small proportion will the unsmooth component cause difficulties. Next we introduce the two classes of problem.

1.2 First problem: bootstrap inference for distributions with both ordinal and continuous components

In some applications it is common to encounter a data distribution that is a mixture of an atom at the origin and a continuous component on the positive

half-line. Examples include the the cost of health care (e.g. Zhou, Melfi and Hui, 1997) and the proportion of an account that an audit determines to be in error (e.g. Cox and Snell, 1979; Azzalini and Hall, 2000). In the second example, both 0 and 1 can be atoms of the sampled distribution. In both examples the mean of the mixture, rather than the mean of just the continuous component, is of interest.

If all the data are ordinal and lie within a relatively narrow range, for example if the costs or proportions in the respectively examples are distributed among only a half-dozen equally-spaced bins, then the lattice nature of the data needs careful attention if bootstrap methods are to be used to construct confidence intervals for the mean. Indeed, particular difficulties associated with this case were addressed in the first detailed theoretical treatment of bootstrap methods for distribution approximation; see Singh (1981). One way of alleviating these difficulties is to use smoothed bootstrap methods; see for example Hall (1987a). On the other hand, no special treatment is required if just the positive part of the sampled distribution is addressed, provided this portion of the distribution is smooth.

This begs the question of what should be done in the mixture case. Does the implicit smoothing provided by the continuous component overcome potential difficulties caused by the ordinal component? How does the answer to this question depend on the proportion of the ordinal component in the mixture? Our results on the effects of smoothing on distribution approximation allow us to answer both these questions; see sections 3.1 and 4.1.

A related problem is that of smoothing a discrete distribution so as to construct a confidence interval for its mean. One approach is to blur each lattice-valued observation over an interval on either side of its actual value; see for example Clark *et al.* (1997, p. 12). For example, if a random variable Y with this distribution takes only integer values, we might replace an observed value $Y = i$ by $i + \epsilon Z$, where $\epsilon > 0$ and Z is symmetric on the interval $[-1, 1]$. How large does ϵ have to be in order to effectively eliminate rounding errors from an approximation to the distribution of the mean of n values of Y? In particular, can we allow ϵ to decrease with sample size, and if so, how fast? Answers will be given in sections 3.1 and 4.1.

Of course, in this second aspect of the first problem it is the mean of Y, not the mean of $X = Y + \epsilon Z$, about which we wish to make inference. However, the mean of ϵZ is known, and so it is a trivial matter to progress from a confidence interval for $E(X)$ to one for $E(Y)$.

1.3 Second problem: confidence bands for distribution functions

Let $\mathcal{U} = \{U_1, \ldots, U_n\}$ denote a random sample drawn from a distribution F, and write $\widehat{F}$ for the empirical distribution function based on $\mathcal{U}$. Then, with $z_{\alpha/2}$ denoting the upper $\frac{1}{2}\alpha$-level point of the standard normal distribution,

$\widehat{F} \pm \{n^{-1}\widehat{F}(1-\widehat{F})\}^{1/2} z_{\alpha/2}$ is a conventional confidence band for F founded on normal approximation, with nominal pointwise coverage $1-\alpha$. In more standard problems, involving a mean of smoothly distributed random variables, the coverage accuracy of such a band would equal $O(n^{-1})$. In the present setting, however, owing to asymmetric rounding errors that arise in approximating the discrete Binomial distribution by a smooth normal one, coverage error of even a two-sided symmetric confidence band is in general no better than $O(n^{-1/2})$.

A particularly simple way of smoothing in this setting, and potentially overcoming difficulties caused by rounding errors, is to use kernel methods. Let K, the kernel, be a bounded, symmetric, compactly supported probability density, write L for the corresponding distribution function, and let h be a bandwidth. Then

$$\widehat{F}h(u) = n^{-1} \sum_{i=1}^{n} L\Big(\frac{u - U_i}{h}\Big) \tag{1.1}$$

is a smoothed kernel estimator of F. We may interpret $E\{\widehat{F}h(u)\}$ in at least two different ways: firstly, as the mean of a sample drawn from a mixture of two distributions, one taking only the values 0 and 1 (the latter with probability $F(u - hc)$, where $[-c, c]$ denotes the support of K), and the other having a smooth distribution (equal to that of $L\{(u - U_i)/h\}$, conditional on $(u - U_i)/h$ lying within the support of K); and secondly, as the distribution function of $X = Y + hZ$, where Y and Z have distribution functions F and L, respectively. Hence, this problem and those described in section 1.2 have identical roots.

The bias of $\widehat{F}h(u)$, as an estimator of $F(u)$, equals $O(h^2)$ provided F is sufficiently smooth. In relative terms its variance differs from that of $\widehat{F}(u)$ by only $O(h)$. See Azzalini (1981), Reiss (1981) and Falk (1983) for discussion of these and related properties. Together these results suggest that taking h as small as possible is desirable, since then h would have least effect on moment properties.

Indeed, the moment properties suggest that $h = O(n^{-1})$ might give the $O(n^{-1})$ coverage error seen in conventional problems. However, it may be shown that this size of bandwidth is not adequate for removing difficulties caused by lack of smoothness of the distribution of $\widehat{F}$. With $h = O(n^{-1})$, rounding errors still contribute terms of order $n^{-1/2}$ to coverage error of two-sided confidence bands. Can we choose h large enough to overcome these problems, and yet small enough to give an order of coverage accuracy close to the "ideal" $O(n^{-1})$? And even if this problem has a theoretical solution, can good coverage accuracy be achieved empirically? These questions will be answered in sections 3.2 and 4.2, where we shall propose and describe an empirical bandwidth-choice method in the confidence band problem. Additionally we shall show that our approach to the problem of smoothed distribution estimation, via sampling from a mixture distribution, leads to particularly simple derivations of Edgeworth expansions.

There is of course an extensive literature of the problem of bandwidth choice for kernel estimation of distribution functions. It includes both plug-in and cross-validation methods; see Mielniczuk, Sarda and Vieu (1989), Sarda (1993), Altman and Legér (1995), and Bowman, Hall and Prvan (1998). However, in all

these cases the bandwidths that are proposed are of asymptotic size $n^{-1/3}$, much larger than n^{-1}. They are appropriate only for estimation of the distribution function curve, not for confidence interval or band construction, and produce relatively high levels of coverage error if used for the latter purpose. The class of distribution and density estimation problems is characterised by an interesting hierarchy of bandwidth sizes: $n^{-1/5}$ for estimating a density curve, $n^{-1/3}$ for distribution curve estimation, and a still smaller size, approximately $n^{-1} \log n$ (as we shall show in section 3.2), for constructing two-sided confidence bands for a distribution function.

2 Distribution-Approximation Difficulties Caused by Lack of Smoothness

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with the distribution of X, and let $\bar{X} = n^{-1} \sum_i X_i$ denote the sample mean. Many explanations for the small-sample performance of bootstrap approximations to the distribution of $\bar{X}$ are based on properties of its Edgeworth expansion. A *formal* expansion exists under moment conditions alone. In particular, provided only that

$$E\big(|X|^{k+2}\big) < \infty\,, \tag{2.1}$$

the formal Edgeworth expansion up to terms in $n^{-k/2}$ is well defined; it is

$$Q_k(x) = \Phi(x) + n^{-1/2}\,\pi_1(x)\,\phi(x) + \ldots + n^{-k/2}\,\pi_k(x)\,\phi(x)\,, \tag{2.2}$$

where Φ and ϕ are standard normal distribution and density functions, respectively, and π_j is a polynomial of degree $3j-1$, odd or even according as j is even or odd respectively, its coefficients depending only on the first $j+2$ moments of X. In particular, $\pi_1(x) = \frac{1}{6}\,\beta\,(1-x^2)$, where $\beta = E(X-EX)^3/(\mathrm{var}X)^{3/2}$. See for example Hall (1992, Chapter 2). These results have straightforward extensions to the Studentised case, which we shall discuss in section 5.5.

If, in addition to the moment assumption (2.1), the distribution of X is smooth (for example if it is absolutely continuous), Q_k can provide an accurate approximation to the standardised distribution of $\bar{X}$. For example, if the distribution of X has a bounded density, and if we define $\mu = E(X)$ and $\sigma^2 = \mathrm{var}(X)$, then

$$P\big\{n^{1/2}(\bar{X}-\mu)/\sigma \le x\big\} = Q_k(x) + o\big(n^{-k/2}\big)\,, \tag{2.3}$$

uniformly in x, as $n \to \infty$. The performance of bootstrap methods rests heavily on this result, through the property that the bootstrap provides a particularly accurate estimate of the term Q_k on the right-hand side of (2.3).

However, (2.3) fails if the sampled distribution is lattice. For example, if $n\bar{X}$ has the Binomial $\mathrm{Bi}(n,q)$ distribution, where $0 < q < 1$, then (2.3) holds only if we add, to the right-hand side, a continuity-correction term for each order from $n^{-1/2}$ to $n^{-k/2}$ inclusive. Such terms compensate for errors introduced by approximating the relatively rough Binomial distribution by a smooth function.

In particular, if the sampling distribution is supported on the set of integers and has lattice span 1, and if we define

$$D_k(x) = Q_k(x) - \sum_{j \le nx} Q_k'\{(j - n\mu)/(n^{1/2}\sigma)\},$$

then the "corrected" form of (2.3) holds:

$$P\{n^{1/2}(\bar{X} - \mu)/\sigma \le x\} = Q_k(x) + D_k(x) + o(n^{-k/2}) \tag{2.4}$$

uniformly in x. See for example pp. 237–241 of Bhattacharya and Rao (1976). Of course, (2.4) has analogues in the case of other lattice distributions. In these general cases we may express $D_k(x)$ as an expansion with terms of size $n^{-j/2}$, for $1 \le j \le k$. The term in $n^{-1/2}$ equals

$$D_{k1}(x) = n^{-1/2}\,\sigma^{-1}\,S(\sigma n^{1/2}x + n\mu)\,\phi(x),$$

where $S(u) = \langle u \rangle - u + \frac{1}{2}$ and $\langle u \rangle$ denotes the integer part of u. The well-known continuity correction, applied for example to normal approximations to the Binomial distribution, adjusts for $D_{k1}(x)$.

We shall show in section 5, however, that if the distribution of X is smoothed through being a mixture of only a small proportion of a continuous distribution, then all aspects of the continuity correction $D_k(x)$ may be dispensed with. That is, $D_k(x)$ may be dropped from (2.4), and (2.3) holds for all $k \ge 1$.

The implications of this result for coverage accuracy of confidence regions can be considerable. To appreciate this point, note that since $\pi_1(x)$ at (2.2) is symmetric in x then, in the case of a smooth sampled distribution, potential coverage errors of size $n^{-1/2}$ cancel from the formula for coverage of the two-sided confidence interval $\bar{X} \pm n^{-1/2}\sigma z_{\alpha/2}$. As a result this interval has coverage error $O(n^{-1})$. However, since the correction term $D_k(x)$ is not symmetric in x then this property fails when the sampled distribution is unsmooth, and there the order of coverage error is only $O(n^{-1/2})$, even for symmetric, two-sided confidence intervals. Moreover, a conventional continuity correction does not remove all the error of this size; taking that approach, the best that can generally be achieved is to produce a conservative confidence interval where the coverage error is dominated by, rather than equal to, the nominal level plus $O(n^{-1})$. See Hall (1982, 1987a) for discussion of this issue.

Of course, these results have direct analogues in the Studentised case; in the discussion above we have treated the non-Studentised case, where σ is assumed known, only for convenience.

3 Overcoming Difficulties Caused by Lack of Smoothness

3.1 Solution to first problem

Suppose the distribution of X is obtained by mixing a smoothly distributed random variable Y (for example, one having a bounded probability density) with an arbitrary but nondegenerate random variable Z, in proportions p and $1-p$ respectively, where p may depend on n. We wish to know the effect that any smoothing conferred by the distribution of Y has on the distribution of a mean $\bar{X}$ of n independent random variables distribution as X.

It will be shown in section 5 that if $n^{-1}\log n = o(p)$ then the discretisation-error term D_k is negligible, and in fact $\sup_x |D_k(x)| = o(n^{-k/2})$. As a result, the distribution of $\bar{X}$ is accurately approximated by its formal Edgeworth expansion, to any order that is permitted by the number of moments enjoyed by the distribution of X. This property applies equally to the distributions of Studentised and non-Studentised means; in both cases, the comparatively small amount of smoothing obtained when $n^{-1}\log n = o(p)$ is nevertheless sufficient to compensate for highly unsmooth features of the other component of the sampling distribution.

We shall also note in section 5 that these results extend to applications of the bootstrap. Indeed, all those properties of the bootstrap that are valid whenever a fixed sampled distribution is accurately approximated by its formal Edgeworth expansion (see e.g. Hall, 1992, Chapter 3), continue to hold for our mixture distribution, provided $n^{-1}\log n = o(p)$.

Of course, these results are somewhat asymptotic in character, although the particularly small lower bound to the effective value of p suggests that in most cases the results will be available in practice. Numerical work in section 4.1 will bear this out. In a specific, practical problem an empirical method for determining whether p is sufficiently large is to explore the problem by Monte Carlo means: model the distribution of the smooth component of the sampled distribution, and, taking the mixing proportion equal to its naive estimate, simulate to ascertain the effect of discretisation error in the context of the model.

In the case of specific component distributions (e.g. a normal smooth component and a Bernoulli lattice component) it can be shown that the constraint $n^{-1}\log n = o(p)$ is necessary as well as sufficient for formal Edgeworth approximation to be valid at all orders. In more general cases it is readily proved that the less stringent constraint $n^{-1} = O(p)$ is not sufficient.

Very similar results may be derived in the related problem of smoothing the distribution of an integer-valued random variable Y by adding to it, rather than mixing it with, a continuous component. That is, we replace Y by $Y + \epsilon Z$, where $\epsilon > 0$ and Z has a continuous distribution. As long as $\epsilon = \epsilon(n)$ decreases to 0 more slowly than $n^{-1}\log n$, this modification allows us to approximate the distribution of the mean of $Y + \epsilon Z$ by its formal Edgeworth expansion to any or-

der; see section 5.3. If the distribution of Z is symmetric then the distributions of both Y and $Y + \epsilon Z$ have the same mean and skewness, and their variances differ only to order ϵ^2. Moreover, the "converse" results described in the previous paragraphs have direct analogues in the setting of additive smoothing of a discrete distribution.

3.2 Solution to second problem

Recall from section 1.3 that we seek a pointwise, $(1 - \alpha)$-level confidence band for the distribution function F. We noted there that the standard normal-approximation band, $\widehat{F} \pm \{n^{-1}\widehat{F}(1 - \widehat{F})\}^{1/2} z_{\alpha/2}$, has only $O(n^{-1/2})$ coverage accuracy, owing to uncorrected discretisation errors. We suggest instead the smoothed band,

$$\widehat{F}h \pm \left\{n^{-1}\widehat{F}h(1 - \widehat{F}h)\right\}^{1/2} z_{\alpha/2} \tag{3.1}$$

where $\widehat{F}h$ is as defined at (1.1). We shall show at the end of this section that by taking $h = n^{-1}(\log n)^{1+\epsilon}$, for any $\epsilon > 0$, coverage error of this band is reduced to $O(h)$. That is only a little worse than the $O(n^{-1})$ level encountered in related problems, where the sampled distribution is smooth.

These properties are highly asymptotic in character, however. To achieve a good level of performance in practice we suggest the following approach. Using standard kernel methods, compute an estimator of the density $f = F'$ based on the sample $\mathcal{U}$. For example, if employing the same kernel K as before, the estimator would be

$$\hat{f}_{h_1}(u) = (nh_1)^{-1} \sum_{i=1}^{n} K\Big(\frac{u - U_i}{h_1}\Big)\,,$$

where h_1 is a bandwidth the size of which is appropriate to density estimation. (In particular, h_1 would generally be computed using either cross-validation or a plug-in rule; it would be of size $n^{-1/5}$, in asymptotic terms.) Let $\widehat{F}_{h_1}(u) = \int_{v \le u} \hat{f}_{h_1}(v)\, dv$ denote the corresponding distribution function. Draw bootstrap resamples $\mathcal{U}^* = \{U_1^*, \ldots, U_n^*\}$ by resampling from the distribution with this density, conditional on $\mathcal{U}$, and use them to compute the bootstrap version $\widehat{F}h^*$ of the smoothed distribution estimator $\widehat{F}h$, this time using bandwidth h rather than h_1. Calculate the corresponding confidence band, $\widehat{F}h^* \pm \{n^{-1}\widehat{F}h^*(1 - \widehat{F}h^*)\}^{1/2} z_{\alpha/2}$, and for each u compute the bootstrap probability $\beta_\alpha(u, h)$ that this band contains $\widehat{F}_{h_1}(u)$:

$$\begin{aligned}\beta_\alpha(u,h) = P\Big(&\widehat{F}h^*(u) - \left[n^{-1}\widehat{F}h^*(u)\left\{1 - \widehat{F}h^*(u)\right\}\right]^{1/2} z_{\alpha/2} \le \widehat{F}_{h_1}(u) \\ &\le \widehat{F}h^*(u) + \left\{n^{-1}\widehat{F}h^*(u)\left\{1 - \widehat{F}h^*(u)\right\}\right]^{1/2} z_{\alpha/2} \,\Big|\, \mathcal{U}\Big)\,.\end{aligned}$$

Choose $h = \hat{h}_\alpha$ to render $\beta_\alpha(u, h)$ as close as possible to α over the interval $\mathcal{I}$ where we wish to construct the final confidence band. For example, we might select $\hat{h}_\alpha$ to minimize $A_\alpha(h)$, where

$$A_\gamma(h) = \int_{\mathcal{I}} \{\beta_\gamma(u, h) - (1 - \gamma)\}^2\, du\,.$$

Our confidence band is that defined at (3.1), but with $h = \hat{h}_\alpha$. If desired, an additional level of calibration can be incorporated by choosing $(\gamma, h) = (\hat{\gamma}, \hat{h})$ simultaneously, to minimise $A_\gamma(h)$, and taking the band to be that at (3.1) but with bandwidth $\hat{h}$ and critical point $z_{\hat{\gamma}/2}$ (instead of $z_{\alpha/2}$).

Finally we outline a derivation of the theoretical properties claimed of the confidence band at (3.1). It will be shown in section 5.4 that if h decreases to 0 at a slower rate than $n^{-1} \log n$, i.e. if

$$n\, h(n)/(\log n) \to \infty\,, \tag{3.2}$$

then the smoothed empirical distribution function estimator $\widehat{F}h$, defined at (1.1), admits a formal Edgeworth expansion of any order $k \geq 1$. That is, if $Q_k = Q_{h,k}$ at (2.2) denotes the formal Edgeworth expansion of $\widehat{F}h(u)$ then the analogue of (2.3) holds for each $k \geq 1$:

$$P\Big(n^{1/2}\,\frac{\widehat{F}h(u) - F_h(u)}{\sigma_h(u)} \leq x\Big) = Q_{h,k}(x) + o\big(n^{-k/2}\big) \tag{3.3}$$

uniformly in x, where

$$F_h(u) = E\{\widehat{F}h(u)\} = \int K(v)\, F(u - hv)\, dv\,,$$

$$\sigma_h(u)^2 = n \operatorname{var}\{\widehat{F}h(u)\} = \int L\Big(\frac{u-v}{h}\Big)^2 f(v)\, dv - F_h(u)^2\,.$$

If F'' exists and is bounded in a neighbourhood of u then $F_h(u) = F(u) + O(h^2)$ and $\sigma_h(u)^2 = F(u)\{1 - F(u)\} + O(h)$. Therefore, provided

$$n^{-1} \log n \ll h = O\big(n^{-1/2}\big)\,, \tag{3.4}$$

(3.3) for $k \geq 2$ implies that

$$P\Big(n^{1/2}\frac{\widehat{F}h(u) - F(u)}{[F(u)\{1 - F(u)\}]^{1/2}} \leq x\Big) = Q_{h,k}(x) + O(h)$$

uniformly in x. It may be shown by Taylor expanding the argument of the probability that this implies

$$P\Big(n^{1/2}\frac{\widehat{F}h(u) - F(u)}{[\widehat{F}h(u)\{1 - \widehat{F}h(u)\}]^{1/2}} \leq x\Big) = Q_{h,k}(x) + O(h)\,. \tag{3.5}$$

Since the bandwidth $h = n^{-1}(\log n)^{1+\epsilon}$ satisfies (3.4) then the claims made immediately below (3.1) follow from (3.5).

Another advantage of our approach is that it leads to particularly simple derivations of detailed Edgeworth expansions. Indeed, once one appreciates that the problem can be posed in terms of sampling from a mixture, (3.3) immediately gives a simple form of the expansion, to arbitrarily high order. Deriving the expansion in more traditional form, with terms of orders $n^{-i/2}h^j$ for $i, j \geq 0$ (rather than simply $n^{-i/2}$), is only a matter of Taylor expanding the quantities σ_h and $Q_{h,k}$ at (3.3). A different argument, based on intrinsic properties of the smoothed distribution function, was given by García-Soidán, González-Manteiga and Prada-Sánchez (1997). In addition to the complexity of that technique, it requires more severe conditions on the smoothness of K.

4 Numerical Properties

4.1 Effects of different mixing proportions in the first problem

We conducted a simulation study to assess the effects of mixing proportions on coverage accuracy of two-sided confidence intervals based on either Studentised or non-Studentised means. We generated 1000 samples of sizes $n = 10$ and 20 from a mixture of a discrete Bernoulli distribution with probability of success 0.1 and different continuous distributions: chi-squared distributions with two, four and six degrees of freedom, and a standard normal distribution. Figure 1 graphs coverage probabilities for two-sided 95% confidence intervals in both Studentised and non-Studentised cases, where the endpoints of the intervals are taken to be $\bar{X} \pm 1.96 n^{-1/2} \sigma$ and $\bar{X} \pm 1.96 n^{-1/2} \hat{\sigma}$, respectively, and $\hat{\sigma}$ is the bootstrap standard deviation. Coverage accuracy in the non-Studentised case is high for even small proportions of continuous data, as argued in section 3.1. More difficulties are experienced in the Studentised case, however. There, increasing the proportion of continuous data has a more marked influence on coverage accuracy. Analogous results are obtained for one-sided confidence intervals, except that there the effect of the proportion of continuous data is confounded with the influence of skewness which now has a significant effect on coverage accuracy for different sample sizes.

4.2 Effect of different mixing proportions in the second problem.

Numerical studies which are not detailed here show that for small bandwidths, before bias becomes a significant problem, coverages of smoothed confidence intervals for distribution functions increase monotonically with increasing bandwidth. This is a consequence of the variability of smoothed distribution estimators decreasing with increasing bandwidth. Confidence intervals usually, although not always, undercover when $h = 0$ and overcover when the bandwidth is taken to equal the value, h_{MSE} say, that gives least mean squared error for a given argument u of the distribution function. As the bandwidth is increased from $h = 0$ to h_{MSE} it typically passes through a value that, when used to construct a smoothed α-level confidence interval for $F(u)$, gives zero coverage error. The bootstrap method suggested in section 3.2 produces an empirical approximation $\hat{h}_\alpha$ to this interval-optimal bandwidth.

Table 1 gives numerical examples of the performance of $\hat{h}_\alpha$. There we took F to be the standard normal distribution function, although results are similar in other cases; only $u = 0$, where the normal density has zero gradient and, consequently, the bias of a distribution estimator equals $O(h^4)$ rather than $O(h^2)$, is atypical. Columns of Table 1 give approximations to the true coverage of confidence intervals (obtained by averaging over 1000 samples, using $B = 1500$ bootstrap simulations) for different values of n. Rows express (a) the confidence interval using the bandwidth $h = h_{\mathrm{MSE}}$ that produces optimal pointwise accu-

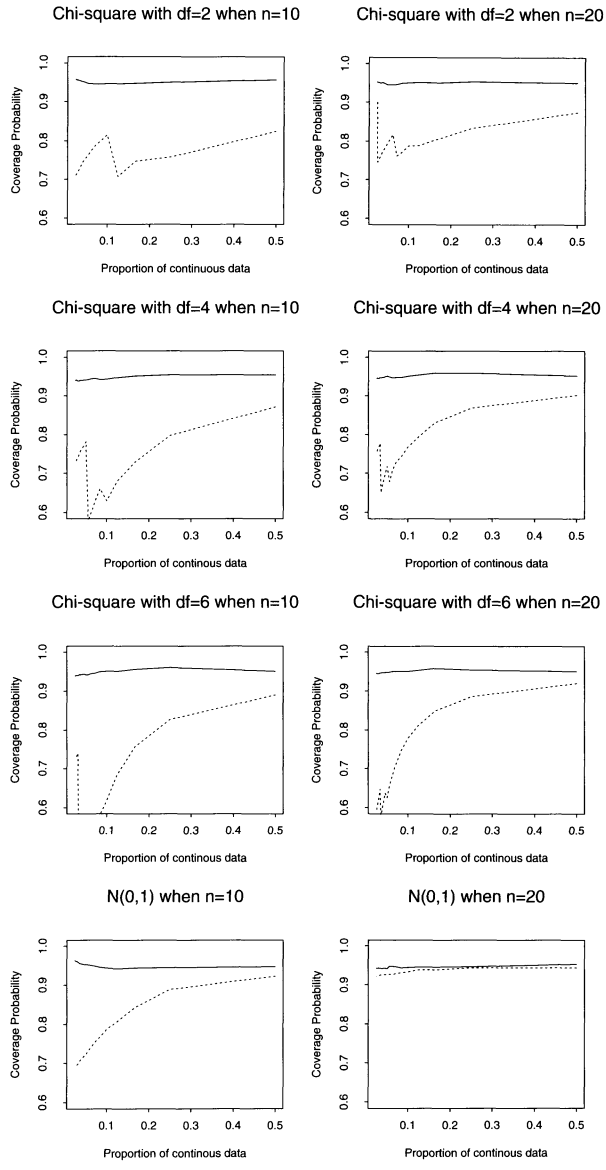

Figure 1 *Coverage probabilities of two-sided 95% confidence intervals.* Solid and dotted lines show coverages of non-Studentised and Studentised intervals, respectively, for the mean of a mixture of a discrete Bernoulli distribution and a chi-squared or a normal distribution.

racy (PTWS); (b) the interval calculated using our bootstrap method (BOOT); and (c) the unsmoothed interval (UNSM). Except when $u = 0$ the coverage for the interval BOOT lies between its counterparts for PTWS and UNSM. In almost every case it is substantially closer to 0.95 than the coverages of either of the other two intervals. In our calculations we employed the distribution version of the Epanechnikov kernel, defined by $L(u) = (3/4)u - (1/4)u^3 + (1/2)$ for $|t| \leq 1$, $L(u) = 0$ if $t < -1$ and $L(u) = 1$ if $t > 1$.

Method	$u = 0.0$		$u = 0.75$		$u = 1.5$	
	n=20	n=50	n=20	n=50	n=20	n=50
BOOT	0.955	0.942	0.941	0.948	0.933	0.954
PTWS	0.990	0.983	0.987	0.986	0.965	0.980
UNSM	0.971	0.918	0.945	0.925	0.766	0.858

N(0,1): the standard normal distribution.
Methods: BOOT, the interval using our bootstrap method;
PTWS, the confidence interval using the bandwidth $h = h_{MSE}$ that produces optimal pointwise accuracy;
UNSM, the unsmoothed interval.

Table 1: *Coverages of different confidence intervals for* $F(u)$. The distribution is standard normal, u denotes the argument at which F is estimated, and rows headed PTWS, BOOT and UNSM represent intervals using the pointwise-optimal bandwidth, the bandwidth $\hat{h}_\alpha$ suggested in section 3.2, and $h = 0$, respectively.

5 Technical Details

5.1 Mixture of discrete and continuous distribution

Let Y be a random variable with the property that its characteristic function $\psi(t) = E(e^{itY})$ satisfies Cramér's condition:

$$\limsup_{|t| \to \infty} |\psi(t)| < 1\,. \tag{5.1}$$

In particular, (5.1) holds if the distribution of Y is absolutely continuous. Let Z denote a random variable independent of Y and having any nondegenerate distribution, and let the distribution of X be a mixture of those of Y and Z in proportions $p : 1 - p$. We shall take p to be a function of sample size, since this allows us to explore the case where $X = Z$ with very high probability. Thus,

$$X = \begin{cases} Y & \text{with probability} p = p(n) \\ Z & \text{with probability} 1 - p\,. \end{cases} \tag{5.2}$$

Given this distribution of X, define the formal Edgeworth expansion Q_k as at (2.2), and put $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$. Note that all moments of X depend on n, through $p(n)$.

Theorem 5.1. *Assume the distribution of Y satisfies* (5.1)*, and that the distribution of X is given by* (5.2)*. Suppose too that the distribution of Z is non-degenerate, that*

$$E(|Y| + |Z|)^{k+2} < \infty \tag{5.3}$$

where $k \geq 1$, and that

$$p(n) \to 0 \quad \textit{and} \quad \lim_{n\to\infty} n\,p(n)/(\log n) = \infty \tag{5.4}$$

as $n \to \infty$. Then

$$P\{n^{1/2}(\bar{X} - \mu)/\sigma \leq x\} = Q_k(x) + o(n^{-k/2}) \tag{5.5}$$

uniformly in x.

Note particularly that (5.4) requires only a very small proportion, not much larger than $O(n^{-1} \log n)$, of the X_i's to be equal to the smoothly distributed Y_i's. Furthermore, the Edgeworth expansion at (5.5) involves no continuity-correction term. Therefore, "a small amount of smoothness goes a long way" in removing any effects of discreteness of the distribution of the sample mean.

5.2 Bootstrap form of Theorem 5.1

Let $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denote a resample drawn by sampling randomly, with replacement, from $\mathcal{X} = \{X_1, \ldots, X_n\}$. Let S^2 be the variance of $\mathcal{X}$ (defined using divisor n rather than $n - 1$), let $\bar{X}^*$ denote the mean of $\mathcal{X}^*$, and let $\widehat{Q}_k$ be the empirical form of Q_k, in which each population moment is replaced by its sample counterpart.

Theorem 5.2. *Assume the conditions of Theorem 5.1. Then*

$$P\{n^{1/2}(\bar{X}^* - \bar{X})/S \leq x | \mathcal{X}\} = \widehat{Q}_k(x) + o_p(n^{-k/2})\,, \tag{5.6}$$

uniformly in x.

The first term in Q_k, of size $n^{-1/2}$, depends on only the first three moments of the distribution of X. Provided $E(|X| + |Y|)^6 < \infty$, these three moments differ from their sample counterparts only by order $n^{-1/2}$. Therefore, taking $k \geq 2$ and subtracting (5.5) and (5.6), we deduce that

$$P\{n^{1/2}(\bar{X}^* - \bar{X})/S \leq x | \mathcal{X}\} - P\{n^{1/2}(\bar{X} - \mu)/\sigma \leq x\} = O_p(n^{-1})\,,$$

uniformly in x. This is the analogue of second-order correctness in the present setting: the bootstrap approximation to the distribution of the sample mean is accurate to order n^{-1}, not simply $n^{-1/2}$ (as in a conventional normal approximation). Note particularly that this has been achieved through only a small amount of smoothing, by mixing a virtually arbitrary Z distribution with only a little more than proportion $O(n^{-1} \log n)$ of the relatively smooth Y distribution.

5.3 Variant of Theorem 5.1 for distribution smoothing

Let Y and Z be independent variables, as discussed in section 5.1, and in place of (5.2) put $X = Y + \epsilon Z$ where $\epsilon = \epsilon(n)$ is nonrandom. For this definition of X let Q_k be the formal Edgeworth expansion as at (2.2).

Theorem 5.3. *Assume the distributions of Y and Z satisfy* (5.3), *that $X = Y + \epsilon(n)\,Z$, and that* (5.4) *holds with $p(n)$ there replaced by $\epsilon(n)$. Then* (5.5) *holds.*

5.4 Application to first and second problems

Application to the first problem is straightforward, provided the distribution of Z is nondegenerate. If the distribution is degenerate and the condition

$$p(n) \text{ is bounded away from } 0 \tag{5.7}$$

fails, then $\sigma = \sigma(n)$ is not bounded away from 0, and this causes difficulties even in interpreting (5.5). In particular, if (5.7) fails then a formal Edgeworth expansion in powers of $n^{-1/2}$ is no longer appropriate; it should instead be in powers of $\{n\,p(n)\}^{-1/2}$.

However, it is straightforward to show that if (5.7) holds then Theorems 5.1 and 5.2 remain valid when the condition that Z has a nondegenerate distribution is removed. Claims made in section 3.1, about properties of confidence intervals and bootstrap methods in the case of the "first problem" (see section 1.2), now follow directly from Theorems 5.1 and 5.2 and their counterparts for the Studentised mean, discussed in section 5.5.

Next we consider allowing the distributions of Y and Z, and hence X, to vary with n. Theorems 5.1 and 5.2 continue to hold in this case, provided (a) the moment condition (5.3) is strengthened to

$$\text{for some } \epsilon > 0\,, \quad \limsup_{n\to\infty} E\{|Y(n)| + |Z(n)|\}^{k+2+\epsilon} < \infty\,, \tag{5.8}$$

(b) the variance of Z is bounded away from 0 in the limit, i.e.

$$\liminf_{n\to\infty} \operatorname{var}\{Z(n)\} > 0\,, \tag{5.9}$$

and (c) the smoothness condition (5.1) holds in a uniform sense, i.e.

$$\limsup_{|t|\to\infty} \sup_{n\ge 1} \bigl|E\bigl[\exp\{it\,Y(n)\}\bigr]\bigr| < 1\,. \tag{5.10}$$

(The analogue of (5.9) for Y follows from (5.10).)

Claims made in section 3.2, about performance of bootstrap methods in the case of the "second problem" (see sections 1.3 and 3.2), follow from Theorems 5.1 and 5.2 under these more general conditions. To appreciate why, note that if the kernel K whose integral equals L is compactly supported, and if the

distribution of the random variable U has a continuous density, then we may interpret $X = L\{(x-U)/h\}$ as being of the form (5.2). In that representation, Y has the distribution of $L\{(x-U)/h\}$ conditional on $x-h < U < x+h$, and Z has a Bernoulli distribution with

$$P(Z=1) = P\{U < x-h \,|\, U \notin (x-h, x+h)\}\,, \quad P(Z=0) = 1 - P(Z=1)\,.$$

(Here we have assumed, without loss of generality, that the support of K equals $[-1,1]$.) If in addition K is bounded and the distribution of U has a bounded density then (5.8)–(5.10) hold, and (5.4) is equivalent to (3.2).

5.5 Further generalisations and extensions

The theorems also apply to the case of the Studentised mean. There we should: (a) alter (5.5) to

$$P\{n^{1/2}(\bar{X}-\mu)/S \le x\} = R_k(x) + o(n^{-k/2})\,,$$

where R_k is the formal Edgeworth expansion corresponding to the Studentised mean; (b) strengthen the moment condition (5.3) to

$$E(|Y|+|Z|)^{2k+4} < \infty\,; \tag{5.11}$$

and (c) change the smoothness assumption (5.1) to

$$\limsup_{|t|+|s|\to\infty} \left|E\{\exp(itY + istY^2)\}\right| < 1\,. \tag{5.12}$$

Alternatively, the original moment condition can be retained but a more restrictive smoothness assumption imposed; compare Hall (1987b). To clarify the differences between the formal Edgeworth expansions Q_k and R_k we note that R_k also admits a formula like (2.2), but with different polynomials π_k. In particular the polynomial π_1 now equals $\frac{1}{6}\beta(2x^2+1)$, instead of $\frac{1}{6}\beta(1-x^2)$. See Hall (1992, Chapter 2) for discussion of these issues.

Likewise, Theorems 5.1 and 5.2 can be extended to the so-called "smooth function model", where $\bar{X}$ is replaced by a smooth function of an r-vector of means. In this case the r-variate versions of (5.11) and (5.12) are sufficient. In each generalisation, condition (5.4) on the mixing proportion may be retained.

Theorems 5.1 and 5.2 also continue to hold if, instead of defining X by (5.2), we take $X_i = Y_i$ for $1 \le i \le \langle np \rangle$, and $X_i = Z_i$ for $\langle np \rangle < i \le n$, where $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ denote independent sequences of independent copies of Y and Z, respectively, where $\langle np \rangle$ denotes the integer part of np. None of the other assumptions needs to be altered; in particular, condition (5.4) on $p = p(n)$ may be retained. However, these variants of the theorems appear to have relatively few statistical applications.

5.6 Outline proof of Theorem 5.1

The derivation is based on characteristic functions and Fourier inversion. It is similar to that in traditional cases (e.g. Petrov, 1975, Chapter 5), with

the exception of the method for bounding the difference, δ say, between the characteristic functions of the left-hand side of (5.5) and of the term Q_k on the right-hand side. Using standard arguments one may obtain the bound $|t^{-1}\delta(t)| \le \xi n^{-k/2} \exp(-\eta t^2)$ for $|t| \le \zeta n^{1/2}$, where $\xi > 0$ can be arbitrarily small and $\eta, \zeta > 0$ depend on ξ but not on n. For $|t| > \zeta n^{1/2}$ one may establish the bound $C_2(1 - C_3)^{np(n)}$, where $C_2 > 0$ and $C_3 \in (0, 1)$ depend on ζ but not on n. Assuming p satisfies (5.4) we may deduce from these bounds, by taking ξ arbitrarily small, that the integral of $|t^{-1}\delta(t)|$ over the interval $(-n^{C_4}, n^{C_4})$, for any $C_4 > 0$, equals $o(n^{-k/2})$, as has to be shown in order to complete the proof.

The proof of Theorem 5.2 is similar, and may be based on arguments of Hall (1992, section 5.2). □

Peter Hall and Xiao-Hua Zhou
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200, Australia

Xiao-Hua Zhou
Division of Biostatistics
Department of Medicine
Indiana University School of Medicine
RG/4th Floor
Regenstrief Health Center
1050 Wishard Boulevard
Indianapolis, IN 46202, USA

Bibliography

[1] Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *J. Statist. Plan. Infer.* **46**, 195–214.

[2] Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68**, 326–328.

[3] Azzalini, A. and Hall, P. (2000). Reducing variability using bootstrap methods with qualitative constraints. *Biometrika*, to appear.

[4] Bhattacharya, R.N. (1967). Berry-Esseen bounds for the multi-dimensional central limit theorem. PhD Dissertation, University of Chicago.

[5] Bhattacharya, R.N. (1968). Berry-Esseen bounds for the multi-dimensional central limit theorem. *Bull. Amer. Math. Soc.* **74**, 285–287.

[6] Bhattacharya, R.N. (1970). Rates of weak convergence for the multidimensional central limit theorem. *Teor. Verojatnost. i Primenen* **15**, 69–85.

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

[8] Bhattacharya, R.N. and Rao, R. Ranga (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

[9] Bowman, A.W., Hall, P. and Prvan, T. (1998). Cross-validation for the smoothing of distribution functions. *Biometrika* **85**, 799–808.

[10] Clark, L.A., Cleveland, W.S., Denby, L. and Liu, C. (1997). Modeling customer survey data. Manuscript.

[11] Cox, D.R. and Snell, E.J. (1979). On sampling and the estimation of rare errors. *Biometrika* **66**, 125–132. Correction *ibid* **69** (1982), 491.

[12] Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statist. Neer.* **37**, 73–83.

[13] García-Soidán, P.H., González-Manteiga, W. and Prada-Sánchez, J.M. (1997). Edgeworth expansions for nonparametric distribution estimation with applications. *J. Statist. Plann. Inf.* **65**, 213–231.

[14] Hall, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika* **69**, 647–652.

[15] Hall, P. (1987a). On the bootstrap and continuity correction. *J. Roy. Statist. Soc.* Ser. B **49**, 82–89.

[16] Hall, P. (1987b). Edgeworth expansion for Student's t-statistic under minimal moment conditions. *Ann. Probab.* **15**, 920–931.

[17] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

[18] Mielniczuk, J., Sarda, P. and Vieu, P. (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plan. Infer.* **23**, 53–69.

[19] Petrov, V.V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

[20] Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.* **8**, 116–119.

[21] Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plan. Infer.* **35**, 65–75.

[22] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.

[23] Zhou, X.H., Melfi, A. and Hui, S.L. (1997). Methods for comparison of cost data. *Ann. Internal Med.* **127**, 752–756.