The Annals of Statistics
1992, Vol. 20, No. 3, 1594–1610

BOOTSTRAP ESTIMATION OF CONDITIONAL DISTRIBUTIONS

BY JAMES BOOTH, PETER HALL AND ANDREW WOOD

Australian National University

Techniques are developed for bootstrap estimation of conditional distributions, with application to confidence intervals and hypothesis tests for one parameter, conditional on the value of an estimator of another. Both Monte Carlo and saddlepoint methods for approximating bootstrap distributions are considered, and empirical methods are suggested for implementing these techniques. For example, in the case of Monte Carlo methods, we suggest empirical techniques for selecting both the smoothing parameter, necessary to define the estimator, and the importance resampling probabilities, required for efficient bootstrap simulation. The smoothing parameter depends critically on the number of Monte Carlo simulations, as well as on the data. Both our theoretical and numerical results indicate that pivoting can substantially improve performance.

1. Introduction. The basics of bootstrap distribution estimation are now quite well understood and have been described particularly well in survey papers by Hinkley (1988) and DiCiccio and Romano (1988). However, very little is known about how bootstrap methods might be used to estimate the conditional bootstrap distribution of the value of one statistic, given a value for the other.

Fisher's "conditionality principle" [e.g., Kendall and Stuart (1979), page 232] for statistical inference is well known. It states, when conducting inference about the first component of a parameter (θ_1, θ_2), based on a statistic (S_1, S_2) where the second component is independent of θ_1, that it suffices to work with the conditional distribution of S_1 given S_2. However, there are other situations where conditional inference is advantageous. In some problems involving paired data, the cost of observing one of the variables is considerably greater than the cost for the other, and the cheaper variable may be principally of interest in shedding light on the more expensive one. For example, this is typically the case when measurements are made of breaking strength and, say, weight of structural members such as timber. The former can usually only be determined by destroying the sample, whereas weight can be measured very inexpensively, and often gives a good indication of breaking strength. If sample sizes are standardized, then a training sample of both variables can be used to provide confidence intervals for the population mean of the more expensive variable, conditional on the observed value of a new sample mean for the cheaper variable. More generally, by using resample sizes different from the training sample size, bootstrap methods may be used to

Received January 1991; revised July 1991.

AMS 1980 *subject classifications*. Primary 62G15; secondary 62G20.

Key words and phrases. Confidence interval, hypothesis test, importance resampling, Monte Carlo, saddlepoint approximation, smoothing parameter.

solve this problem even when the training sample size differs from subsequent sample sizes.

Our aim in this paper is to develop methodology for bootstrap estimation of conditional distributions. We treat the case where a confidence interval or hypothesis test is derived for the value of one parameter, given the value taken by an estimate of another parameter. Other situations, for example that where the training sample and subsequent samples are of different sizes, may be handled similarly. We suggest empirical rules for selecting the amount of smoothing, and the method of resampling, so as to obtain accurate bootstrap estimators of distributions. Particular attention is paid to the issue of pivoting, which is particularly interesting in the present context; here, pivoting demands standardization for correlation as well as variance. In the case where the statistics of interest are means, we describe both Monte Carlo and saddlepoint methods for approximating bootstrap distributions, and compare these two approximations with a normal approximation.

Section 2 introduces a variety of bootstrap methods for approximating conditional distributions. Section 3 describes an illustrative example, Section 4 discusses saddlepoint methods and Section 5 summarizes large-sample theory for Sections 2 and 3. Numerical examples are given in both Sections 3 and 4.

2. Methodology.

2.1. *Introduction and summary.* We develop a systematic approach to the construction of resampling approximations to conditional probabilities. Section 2.2 notes that there are difficulties defining bootstrap estimators directly and suggests that these be overcome by smoothing. Kernel methods are proposed with the degree of smoothing governed by a bandwidth, h. Direct calculation of these estimators would usually be impossible, and so approximate methods based on Monte Carlo simulation should be considered. Section 2.3 develops this approach in the case of nonpivotal percentile-type estimators and introduces the notion of importance resampling. Choice of optimal importance resampling probabilities is addressed in Section 2.5.

In Section 2.4 we argue in favour of pivotal methods, pointing out that for appropriate choices of the bandwidth and the number of simulations, pivotal methods can have an accuracy of $O(n^{-1})$ rather than $O(n^{-1/2})$. A detailed justification of this claim will be given in Section 5. In the context of estimating conditional distributions, pivoting involves standardizing by an estimate of the correlation coefficient as well as Studentizing by an estimate of the sample standard deviation. Versions of these estimates have to be calculated for each resample, and pivotal methods cannot be expected to perform well if stable estimates are not available. Instability typically occurs when the statistic of interest is formed from a ratio, such as the correlation coefficient or a ratio of two means, and the denominator assumes values close to 0. In such circumstances, the simpler, nonpivotal methods described in Section 2.3 would be preferred.

Both pivotal and nonpivotal methods depend on choice of the bandwidth, h, and the number of Monte Carlo resamples, B. For optimal performance, h must depend on B. This is an unusual aspect of the problem, since the idealized bootstrap estimators defined in Section 2.2 depend on h but not on B. It is only when considering practical calculation of those estimators that we must introduce Monte Carlo simulations, and hence B. Section 2.6 suggests practical rules for choosing h as a function of B.

Finally, Section 2.7 describes application of these techniques to the problem of constructing a confidence interval for the true value of one parameter, conditional on the observed value of an estimator of another. A numerical example will be given in Section 3.

It is helpful to summarize our notation. We suppose that there are two unknown parameters θ_1, θ_2, with estimates $\hat{\theta}_1, \hat{\theta}_2$ having asymptotic variances σ_1^2, σ_2^2, and that the latter have estimates $\hat{\sigma}_1^2, \hat{\sigma}_2^2$. It is assumed that $(\hat{\theta}_1 - \theta_1)/\sigma_1$ and $(\hat{\theta}_2 - \theta_2)/\sigma_2$ have an asymptotic joint normal $N(0, 0; 1, 1; \rho)$ distribution. We write $\hat{\rho}$ for a sample estimate of ρ. If estimates such as $\hat{\theta}_1, \hat{\theta}_2, \dots$ are calculated for a resample rather than the original sample, then this fact is indicated by an asterisk * if the resample was drawn uniformly from the sample, and by a dagger † if the resample was drawn by importance resampling.

It is assumed that the number of resamples, B, is only of algebraic order in the sample size, n. That is, $B = O(n^c)$ for some $c > 0$. For example, $B \sim e^{nc}$ is not allowed. The bandwidth formula which we given in Section 2.6 guarantees that if B is no larger than algebraic, then h is no smaller than algebraic, meaning that $h^{-1} = O(n^c)$ for some c. This is important to both the theory in Section 5 and the practical construction of the estimators and their Monte Carlo approximation. In particular, the bootstrap estimates introduced in Section 2.2 will perform erratically if h is geometrically small.

2.2. *Definition of a bootstrap estimate.* Let $\hat{\theta}_1$ and $\hat{\theta}_2$ represent estimates of parameters θ_1 and θ_2, computed from a sample $\mathscr{X}$. Write $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ for the values assumed by $\hat{\theta}_1$ and $\hat{\theta}_2$ when the sample is changed to a resample $\mathscr{X}^*$ drawn randomly, with replacement, from $\mathscr{X}$. Let Pr′ and E' denote probability and expectation, respectively, conditional on $\mathscr{X}$. Formally, we might define

$$\tilde{q} = \Pr'\left(\hat{\theta}_1^{*'} \le x \middle| \hat{\theta}_2^* = y\right) \tag{2.1}$$

to be the "naive" bootstrap estimate of $q = \Pr(\hat{\theta}_1 \le x | \hat{\theta}_2 = y)$. However, there are both practical and theoretical difficulties with this proposal. Even if y happens to be an atom of the bootstrap distribution, the value of $\tilde{q}$ may bear little relation to q. For example, if the data come from a continuous distribution and we take $y = \hat{\theta}_2$, then it will typically be the case that with probability 1, $\hat{\theta}_2^* = \hat{\theta}_2$ if and only if the resample $\mathscr{X}^*$ is identical to the sample $\mathscr{X}$. A case in point is that where $\hat{\theta}_2$ denotes the sample mean. In this circumstance, the value of $\tilde{q}$ at (2.1) reduces to 1 if $\hat{\theta}_1 \le x$ and to 0 otherwise. That is, the "naive" bootstrap estimate of $\Pr(\hat\theta_1 \le x | \hat\theta_2 = \hat\theta_{2\text{obs}})$ is just the indicator function of the event $\hat\theta_1 \le x$, which is quite unsatisfactory.

A more appropriate estimate would be one which averaged over values $\hat\theta_2^*$ in a neighbourhood of y. For example, we might take

$$\tilde q = \Pr'\left(\hat\theta_1^* \le x | y - h \le \hat\theta_2^* \le y + h\right) \tag{2.2}$$

for an appropriate small number h.

To generalize the estimate at (2.2), let K denote a kernel function of the type familiar in problems on nonparametric curve estimation [e.g., Silverman (1986), Chapter 3, and Härdle (1990), Section 3.1]. In particular, we ask that $\int K = 1$ and $\int yK(y)\,dy = 0$. Put

$$\begin{aligned}\hat\mu_1 &= h^{-1}E'\left[I\left(\hat\theta_1^* \le x\right)K\left\{\left(y - \hat\theta_2^*\right)/h\right\}\right],\\ \hat\mu_2 &= h^{-1}E'\left[K\left\{\left(y - \hat\theta_2^*\right)/h\right\}\right],\\ \hat q &= \hat\mu_1/\hat\mu_2.\end{aligned} \tag{2.3}$$

If we take $K(u) = I(|u| \le 1)/2$, then the estimates at (2.2) and (2.3) agree precisely.

2.3. *Monte Carlo approximation.* The kernel estimate at (2.3) may be approximated by Monte Carlo simulation, leading to an estimate similar to that employed in problems of nonparametric regression. Let $\mathscr{X}_1^*, \dots, \mathscr{X}_B^*$ denote independent resamples drawn randomly, with replacement, from the sample $\mathscr{X} = \{X_1, \dots, X_n\}$. Let $(\hat\theta_{1i}^*, \hat\theta_{2i}^*)$ denote the version of $(\hat\theta_1, \hat\theta_2)$ computed from $\mathscr{X}_i^*$. Then

$$\hat\mu_1^* = \frac{1}{Bh}\sum_{i=1}^{B} I\left(\hat\theta_{1i}^* \le x\right)K\left\{\frac{y - \hat\theta_{2i}^*}{h}\right\},$$

$$\hat\mu_2^* = \frac{1}{Bh}\sum_{i=1}^{B} K\left\{\frac{y - \hat\theta_{2i}^*}{h}\right\}$$

represent unbiased approximations to $\hat\mu_1, \hat\mu_2$, respectively, in the sense that $E'(\hat\mu_k^*) = \hat\mu_k$ for $k = 1, 2$. This suggests that we take $\hat q^* = \hat\mu_1^*/\hat\mu_2^*$ as a Monte Carlo approximation to $\hat q$.

More generally, the Monte Carlo simulation might be done by importance resampling, as follows. Consider the probability distribution which ascribes mass π_j to the sample value X_j for $1 \le j \le n$, where $\sum \pi_j = 1$. Let $\mathscr{X}_1^\dagger, \dots, \mathscr{X}_B^\dagger$ denote independent resamples drawn randomly according to this rule, and write $M_{ij}^\dagger$ for the number of times X_j appears in $\mathscr{X}_i^\dagger$. Let $(\hat\theta_{1i}^\dagger, \hat\theta_{2i}^\dagger)$ denote the version of $(\hat{\theta}_1, \hat{\theta}_2)$ computed from $\mathscr{X}_i^\dagger$, and put

$$\hat{\mu}_1^\dagger = \frac{1}{Bh} \sum_{i=1}^{B} I(\hat{\theta}_{1i}^\dagger \le x) K\left\{\frac{y - \hat{\theta}_{2i}^\dagger}{h}\right\} \prod_{j=1}^{n} (n\pi_j)^{-M_{ij}^\dagger},$$

$$\hat{\mu}_2^\dagger = \frac{1}{Bh} \sum_{i=1}^{B} K\left\{\frac{y - \hat{\theta}_{2i}^\dagger}{h}\right\} \prod_{j=1}^{n} (n\pi_j)^{-M_{ij}^\dagger}.$$

Then $\hat{\mu}_k^\dagger$ is an unbiased approximation to $\hat{\mu}_k$ for $k = 1, 2$, and so $\hat{q}^\dagger = \hat{\mu}_1^\dagger / \hat{\mu}_2^\dagger$ is a potential Monte Carlo approximation. Of course, we should choose $\pi_1, \dots, \pi_n$ so as to minimize the error in this approximation. Section 2.5 will address this problem.

2.4. *Pivoting.* The issue of pivoting, or Studentizing, has received considerable attention in work on the bootstrap [e.g., DiCiccio and Romano (1988) and Hall (1988)]. In problems where a "stable" estimate $\hat{\sigma}_k^2$ of the variance of $\hat{\theta}_k$ is available, there are advantages in approximating the distribution of $(\hat{\theta}_k - \theta_k)/\hat{\sigma}_k$ rather than that of $\hat{\theta}_k$. These advantages persist in the problem of estimating conditional distributions, as we shall show in Section 5.1. In particular, if $\hat{\rho}$ is a sample estimate of the asymptotic coefficient of correlation, ρ, between $\hat{\theta}_1$ and $\hat{\theta}_2$, then a bootstrap approximation to

$$\begin{aligned} p &= p(u, w) \\ &= \Pr\left[(1 - \hat{\rho}^2)^{-1/2}\left\{(\hat{\theta}_1 - \theta_1)\hat{\sigma}_1^{-1} - \hat{\rho} w\right\} \le u \,\middle|\, (\hat{\theta}_2 - \theta_2)/\hat{\sigma}_2 = w\right] \end{aligned} \tag{2.4}$$

will typically be in error by only $O(n^{-1})$, whereas a bootstrap approximation to $q = \Pr(\hat{\theta}_2 \le x | \hat{\theta}_2 = y)$ will usually be in error by terms of size $n^{-1/2}$. The form of standardization in (2.4) derives from the fact that if (V, W) is approximately normal $N(0, 0; 1, 1, \rho)$, then conditional on $W = w$, $(1 - \rho^2)^{-1/2}(V - \rho w)$ is approximately normal $N(0, 1)$.

To develop a bootstrap approximation to p based on importance resampling, let $\hat{\sigma}_{ki}^\dagger, \hat{\rho}_i^\dagger$ denote the versions of $\hat{\sigma}_k$ and $\hat{\rho}$ computed from the resample $\mathscr{X}_i^\dagger$, and put $V = (\hat{\theta}_1 - \theta_1)/\hat{\sigma}_1$, $W = (\hat{\theta}_2 - \theta_2)/\hat{\sigma}_2$, $U = (1 - \hat{\rho}^2)^{-1/2}(V - \hat{\rho} w)$, $V_i^\dagger = (\hat{\theta}_{1i}^\dagger - \hat{\theta}_1)/\hat{\sigma}_{1i}^\dagger$, $W_i^\dagger = (\hat{\theta}_{2i}^\dagger - \hat{\theta}_2)/\hat{\sigma}_{2i}^\dagger$, $U_i^\dagger = (1 - \hat{\rho}_i^{\dagger 2})^{-1/2}(V_i^\dagger - \hat{\rho}_i^\dagger w)$ and

$$\hat{\lambda}_1^\dagger = \hat{\lambda}_1^\dagger(u, w) = \frac{1}{Bh} \sum_{i=1}^{B} I(U_i^\dagger \le u) K\left\{\frac{w - W_i^\dagger}{h}\right\} \prod_{j=1}^{n} (n\pi_j)^{-M_{ij}^\dagger}, \tag{2.5}$$

$$\hat{\lambda}_2^\dagger = \hat{\lambda}_2^\dagger(w) = \frac{1}{Bh} \sum_{i=1}^{B} K\left\{\frac{w - W_i^\dagger}{h}\right\} \prod_{j=1}^{n} (n\pi_j)^{-M_{ij}^\dagger}. \tag{2.6}$$

Then the desired approximation is $\hat{p}^\dagger = \hat{p}^\dagger(u, w) = \hat{\lambda}_1^\dagger / \hat{\lambda}_2^\dagger$. We also define

$$p_0(u) = p(u, 0), \qquad \hat{p}_0^\dagger(u) = \hat{p}^\dagger(u, 0) = \hat{\lambda}_1^\dagger(u, 0)/\hat{\lambda}_2^\dagger(0). \tag{2.7}$$

In respect of $\hat{\theta}_2$, although not for $\hat{\theta}_1$, the issue of pivoting is often vacuous. This is because we generally wish to compute the probability that $(\hat{\theta}_1 - \theta)/$

$\hat{\sigma}_1 \leq x$, conditional on the *observed* value of $\hat{\theta}_2$, which we denote by $\hat{\theta}_{2\mathrm{obs}}$. In calculating the bootstrap approximation to this probability, we effectively condition on the event that $\hat{\theta}_2^*$ lies within a neighbourhood of $\hat{\theta}_2$, or equivalently, that $(\hat{\theta}_2^* - \hat{\theta}_2)/\hat{\sigma}_2^*$ lies within a neighbourhood of 0. Irrespective of how the Studentizing of $\hat{\theta}_2^*$ is performed, or indeed whether Studentizing is carried out or not, we are effectively conditioning on the event that $\hat{\theta}_2^* - \hat{\theta}_2$ is close to 0. This event is determined by choice of the bandwidth h. Therefore, when the probability being calculated is conditioned on the observed value of $\hat{\theta}_2$, Studentizing of $\hat{\theta}_2$ has a bearing on the scale of the bandwidth h but hardly at all on the accuracy of the bootstrap approximation.

2.5. *Choice of resampling probabilities for importance resampling.* We assume that the sample values in $\mathscr{X} = \{X_1, \dots, X_n\}$ are d-variate and that $\theta_k = g_k(\mu)$ for $k = 1, 2$ are smooth functions of the population mean $\mu = E(X)$. Let $\bar{X} = n^{-1}\Sigma X_j$, put $\hat{\theta}_k = g_k(\bar{X})$, let $z^{(l)}$ denote the lth element of the d-vector z and define

$$g_{kl}(z) = (\partial/\partial z^{(l)})g_k(z), \qquad D_{kj} = \sum_{l=1}^{d} \left(X_j - \bar{X}\right)^{(l)} g_{kl}(\bar{X}), \qquad \hat{s}_k^2 = \sum_{j=1}^{n} D_{kj}^2$$

and $\varepsilon_{kj} = \hat{s}_k^{-1} D_{kj}$. In this setting, $\hat{\sigma}_k^2 = n^{-2}\hat{s}_k^2$ is an estimator of the variance of $\hat{\theta}_k$. Note too that $\Sigma_j \varepsilon_{kj}^2 = 1$ and that $\hat{\rho} = \Sigma \varepsilon_{1j}\varepsilon_{2j}$ estimates the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$.

Let U, W be as in Section 2.4 and define $Y = (1 - \hat{\rho}^2)^{-1/2}\hat{\sigma}_1^{-1}(\hat{\theta}_1 - \theta_1)$. Suppose we wish to estimate $p = \Pr(U \leq u | W = w)$, using the approximant $\hat{p}^{\dagger}$ defined in Section 2.4; or to estimate $p_0 = \Pr(Y \leq u | \hat{\theta} = \hat{\theta}_{2\mathrm{obs}})$, using the approximant $\hat{p}_0^{\dagger}$; or to estimate $q = \Pr(\hat{\theta}_1 \leq x | \hat{\theta}_2 = y)$, using the approximant $\hat{q}^{\dagger}$ introduced in Section 2.3. In the case of $\hat{p}_0^{\dagger}$, take $w = 0$ in the work which follows, and in the case of $\hat{q}^{\dagger}$, take $w = (y - \hat{\theta}_2)/\hat{\sigma}_2$ and $u = (1 - \hat{\rho}^2)^{-1/2}\{(x - \hat{\theta}_1)\hat{\sigma}_1^{-1} - \hat{\rho} w\}$. Let Φ denote the standard normal distribution function. It will be shown in the Appendix that the asymptotically optimal choice of the resampling probabilities π_j is

$$\pi_j = \exp\left\{-(A_1\varepsilon_{1j} + A_2\varepsilon_{2j}) + C\right\}, \tag{2.8}$$

where C is chosen to ensure that $\Sigma \pi_j = 1$, and A_1, A_2 are chosen to minimize

$$\begin{aligned} &\left[\{1 - 2\Phi(u)\}\Phi\left\{u - A_1(1 - \rho^2)^{1/2}\right\} + \Phi(u)^2\right] \\ &\quad \times \exp\left\{(A_1\rho + A_2)w + A_1^2(1 - \rho^2) + \tfrac{1}{2}(A_1\rho + A_2)^2\right\}. \end{aligned} \tag{2.9}$$

If $w = 0$, then this reduces to taking $A_2 = -A_1\rho$ and choosing A_1 to minimize

$$\left[\{1 - 2\Phi(u)\}\Phi\left\{u - A_1(1 - \rho^2)^{1/2}\right\} + \Phi(u)^2\right]\exp\left\{A_1^2(1 - \rho^2)\right\}. \tag{2.10}$$

In (2.9) and (2.10), ρ may be replaced by its sample estimate $\hat{\rho}$ without affecting the asymptotic optimality of these choices of A_1 and A_2.

TABLE 1

The value of $t(u)$ which minimizes $b(t, u)$ in (2.11) for a range of values of u

u	$t(u)$	u	$t(u)$	u	$t(u)$	u	$t(u)$	u	$t(u)$
0.1	−0.06	0.7	−0.43	1.3	−0.76	1.9	−1.05	2.5	−1.30
0.2	−0.13	0.8	−0.49	1.4	−0.81	2.0	−1.09	2.6	−1.35
0.3	−0.19	0.9	−0.55	1.5	−0.86	2.1	−1.13	2.7	−1.39
0.4	−0.25	1.0	−0.60	1.6	−0.91	2.2	−1.18	2.8	−1.43
0.5	−0.31	1.1	−0.66	1.7	−0.96	2.3	−1.22	2.9	−1.47
0.6	−0.37	1.2	−0.71	1.8	−1.00	2.4	−1.26	3.0	−1.51

Note: For nonpositive values of u, use $t(-u) = -t(u)$.

Table 1 lists values of $t = t(u)$ which minimize

$$b(t,u) = \left[\{1 - 2\Phi(u)\}\Phi(u - t) + \Phi(u)^2\right]e^{t^2} \tag{2.11}$$

for selected u's. An empirical approximation to the value of A_1 which minimizes (2.10) is given by $A_1 = (1 - \hat{\rho}^2)^{-1/2}t(u)$. Note that $b(t, u) = b(-t, -u)$, and so $t(-u) = -t(u)$. This explains the symmetry of the efficiency curve $e(u)$ specified in (2.12) and contrasts markedly with the pronounced asymmetry found in the case of importance sampling estimates of unconditional probabilities. See Johns (1988) and Davison (1988) for the latter.

We show in the Appendix that the quantities at (2.9) and (2.10) are asymptotically proportional to the conditional variance of the estimator obtained by importance resampling, and should be replaced by $\Phi(u)\{1 - \Phi(u)\}$ in the case of uniform resampling. The asymptotic efficiency of importance resampling relative to naive uniform resampling is therefore given by

$$e(u) = \frac{\Phi(u)\{1 - \Phi(u)\}}{\min_t\left(\left[\{1 - 2\Phi(u)\}\Phi(u - t) + \Phi(u)^2\right]e^{t^2}\right)}. \tag{2.12}$$

The graph of this function is symmetric about $u = 0$ and cup-shaped, asymptoting rapidly to $+\infty$ as $|u| \to \infty$. For example, $e(1.645) = 2.85$ and $e(1.96) = 4.33$. This indicates that an improvement in efficiency by at least a factor of 2 is achievable in many situations of practical interest, by using our importance sampling procedures.

We should remind the reader that, owing to the smoothing used to construct $\hat{p}_0^\dagger$, this approximant can be biased as well as having error about the mean. Our use of the term "efficiency" above pertains only to a comparison of variance, not mean squared error.

2.6. *Choice of bandwidth.* The key to a simple solution to the bandwidth problem is to observe that, since $(\hat{\theta}_1, \hat{\theta}_2)$ is asymptotically normally distributed, unknown distributions which appear in formulae for optimal bandwidths may be replaced by their normal approximations. We shall illustrate this technique by treating the pivotal case, when conditioning is on the observed value of $\hat{\theta}_2$. That is, we suggest a formula for the bandwidth h used to compute $\hat{p}_0^\dagger$, defined at (2.7). This quantity represents a bootstrap approximation to $p_0 = \Pr\{(1-\hat\rho^2)^{-1/2}\hat\sigma_1^{-1}(\hat\theta_1-\theta_1) \le u | \hat\theta_2 = \hat\theta_{2\mathrm{obs}}\}$.

We assume that the kernel K is of second order, meaning that $\int K = 1$, $\int yK(y)\,dy = 0$, $2\kappa_1 = \int y^2 K(y)\,dy \ne 0$. Arguments in Section 5 and the Appendix produce formulae for the error about the mean, and the conditional bias, of the bootstrap approximant $\hat p^\dagger$. Thus, defining $\hat p_0$ to equal the ratio of the conditional expected values of the numerator and denominator in formula (2.7) for $\hat p_0^\dagger$, we may prove as in the Appendix and Section 5, respectively, that

$$(2.13)\qquad \hat p_0^\dagger = \hat p_0 + (Bh)^{-1/2}\{\kappa_2\beta/\phi(0)\}^{1/2}Z + o_p\{(Bh)^{-1/2}\},$$

$$(2.14)\quad \hat p_0 = \check p_1(u,0) - h^2\kappa_1\rho^2(1-\rho^2)^{-1}u\phi(u) + O_p(h^3) + o_p(n^{-1}).$$

In these formulae, $\kappa_2 = \int K^2$, $\phi = \Phi'$, β is given by (2.10) in the case of importance resampling and by $\beta = \Phi(u)\{1-\Phi(u)\}$ in the case of uniform resampling, the random variable Z is asymptotically distributed as normal $N(0,1)$, and $\check p_1(u,0)$ denotes an Edgeworth approximation to p_0. It is defined as the standard two-term Edgeworth expansion of p_0, up to and including terms of size $n^{-1/2}$, except that population moments are replaced by sample moments [see (5.2)]. Thus, $\check p_1$ does not depend on h, and $\check p_1 - p_0 = O_p(n^{-1})$; see Sections 2.4 and 5.1.

If we regard $\check p_1$ as the target of the bootstrap approximant $\hat p_0^\dagger$, then it follows from (2.13) and (2.14) that the asymptotic mean squared error is given by

$$(Bh)^{-1}\kappa_2\beta\phi(0)^{-1} + h^4\left\{\kappa_1\rho^2(1-\rho^2)^{-1}u\phi(u)\right\}^2.$$

This quantity is minimized by

$$h = \left[\{4B\phi(0)\}^{-1}\kappa_2\beta\right]^{1/5}\left\{\kappa_1\rho^2(1-\rho^2)^{-1}|u|\phi(u)\right\}^{-2/5},$$

which suggests the empirical bandwidth

$$h = \left[\{4B\phi(0)\}^{-1}\kappa_2\hat\beta\right]^{1/5}\left\{\kappa_1\hat\rho^2(1-\hat\rho^2)^{-1}|u|\phi(u)\right\}^{-2/5}.$$

In the latter expression, $\hat\beta$ would be obtained from the formula (2.10) for β by replacing ρ by $\hat\rho$.

For nonzero w, $\phi(0)$ in (2.13) should be replaced by $\phi(w)$, $\check p_1(u,0)$ by $\check p_1(u,w)$ in (5.2), and $\psi(u,0) = \rho^2(1-\rho^2)^{-1}u\phi(u)$ should be replaced by $\psi(u,w)$ defined below (5.5). The corresponding empirical bandwidth is given by

$$h = \left[\{4B\phi(w)\}^{-1}\kappa_2\hat\beta\right]^{1/5}\{\kappa_1|\psi(u,w)|\}^{-2/5},$$

where $\hat\rho$ replaces ρ in $\psi(u,w)$ and $\hat\beta$, and $\hat\beta$ is now given by (2.9). In the nonpivotal case, the situation is somewhat different because of the presence of the unknown parameter ρ in the leading term of the Edgeworth expansion of $q = \Pr(\hat\theta_1 \le x | \hat\theta_2 = y)$; see the discussion at the end of Section 5.1.

2.7. *Construction of confidence intervals.* Suppose we wish to construct a confidence interval for θ_1, conditional on the observed value of $\hat\theta_2$. We shall describe a solution to this problem based on the bootstrap approximant $\hat p_0^\dagger$, defined at (2.7). Methods which utilize nonpivotal statistics, such as those discussed in Section 2.3, may be developed similarly. However, since nonpivotal techniques are usually only first-order correct, meaning that they approximate probabilities with errors of size $n^{-1/2}$ rather than n^{-1}, then they will not usually perform as well as pivotal methods in problems where stable estimates of variance and correlation are available.

Recall that $\hat p_0^\dagger(u)$ approximates $p_0(u) = \Pr\{(1-\hat\rho^2)^{-1/2}\hat\sigma_1^{-1}(\hat\theta_1 - \theta_1) \le u | \hat\theta_2 = \hat\theta_{2\text{obs}}\}$. Given $0 < \alpha < 1$, put

$$\hat u^\dagger_\alpha = \left(\hat p_0^\dagger\right)^{-1}(\alpha) = \inf\left\{u\colon \hat p_0^\dagger(u) \ge \alpha\right\}.$$

Then $\hat u^\dagger_\alpha$ is a bootstrap approximation to the value u_α such that

$$\Pr\left\{\left(1-\hat\rho^2\right)^{-1/2}\hat\sigma_1^{-1}\left(\hat\theta_1 - \theta_1\right) \le u_\alpha \middle| \hat\theta_2 = \hat\theta_{2obs}\right\} = \alpha.$$

Therefore, an approximate one-sided α-level confidence interval for θ_1, conditional on $\hat\theta_2$, is

$$\left(\hat\theta_1 - \left(1-\hat\rho^2\right)^{1/2}\hat\sigma_1\hat u^\dagger_\alpha, \infty\right).$$

An approximate two-sided α-level interval is

$$\left(\hat\theta_1 - \left(1-\hat\rho^2\right)^{1/2}\hat\sigma_1\hat u^\dagger_{(1+\alpha)/2}, \hat\theta_1 - \left(1-\hat\rho^2\right)^{1/2}\hat\sigma_1\hat u^\dagger_{(1-\alpha)/2}\right).$$

3. Application to bivariate means.

A simulation study to examine our approach in the case of bivariate means was conducted. The results, which are summarized in Table 2, were obtained as described below.

One thousand samples of 50 bivariate observations were generated. The generic sample $\mathscr{X} = \{(X_1^{(1)}, X_1^{(2)}), \dots, (X_n^{(1)}, X_n^{(2)})\}$ consisted of independent exponential variates $X_i^{(2)}$, $i = 1, \dots, n$, with common theoretical mean 1, and

TABLE 2
Coverage properties of the pivotal bootstrap, saddlepoint and normal methods

Sample size	$N = 10$		$N = 20$		$N = 30$		$N = 40$		$N = 50$	
	L	U	L	U	L	U	L	U	L	U
True value	5.0%	95.0%	5.0%	95.0%	5.0%	95.0%	5.0%	95.0%	5.0%	95.0%
Method										
B(200)	12.4%	99.0%	7.6%	98.8%	5.5%	98.1%	5.5%	97.2%	5.8%	97.5%
B(500)	12.3%	98.9%	7.5%	98.8%	4.8%	98.1%	5.1%	97.2%	5.8%	97.7%
B(1000)	12.8%	98.7%	7.8%	98.9%	4.6%	98.0%	5.2%	97.3%	5.7%	97.4%
SP	29.9%	97.1%	18.5%	98.4%	15.4%	98.0%	14.0%	97.0%	14.0%	97.7%
Normal	26.3%	99.2%	16.5%	99.1%	14.3%	98.6%	13.0%	97.5%	12.6%	98.1%

Note: Columns labelled L (U) give the proportion of times that the true parameter value was smaller than the lower (upper) confidence limit.

their logarithms $X_i^{(1)} = \log(X_i^{(2)})$, $i = 1, \dots, n$. Our objective was to obtain, conditional on $\hat{\theta}_2 = \sum X_i^{(2)}/n$, a nonparametric confidence interval for the parameter

$$\theta_1 = E(X^{(1)}) = \int_0^\infty \log(x) e^{-x}\, dx = -0.57721\dots .$$

Using the fact that $(D_1, \dots, D_n) = (X_1^{(2)}/\hat{\theta}_2, \dots, X_n^{(2)}/\hat{\theta}_2)$ is independent of $\hat{\theta}_2$ and has a uniform distribution on the simplex

$$\sum_{i=1}^{n} D_i = n, \qquad D_i \geq 0, \qquad i = 1, \dots, n,$$

it is seen that

$$\hat{\theta}_1 | \hat{\theta}_2 =_d n^{-1} \sum \log(D_i) + \log(\hat{\theta}_2).$$

Of course, $\hat{\theta}_1 = \sum X_i^{(1)}/n$. In the simulation study we chose $\hat{\theta}_2 = E(\hat{\theta}_2) = 1$.

For each of the 1000 simulated samples, the calculations outlined in Section 2.7 were performed and two-sided confidence intervals, of nominal level 0.9 and symmetric on the probability scale, were constructed. Three choices of B, the number of bootstrap resamples, were considered: $B = 200$, 500 and 1000. The reduced sample sizes $n = 10$, 20, 30 and 40 in Table 2 were obtained by selecting the first 10, 20, 30 and 40 observations respectively from the size 50 samples.

The results in Table 2 indicate that the pivotal bootstrap approach is insensitive to the choice of B in the range $200 \leq B \leq 1000$. The differences in coverage "on the left" and "on the right" are due to the fact that the underlying distribution is highly skewed. The bootstrap does well on the left when $n \geq 30$, but on the right the bootstrap confidence intervals tend to be too long although there is improvement with increasing n. The performance of the saddlepoint method which is described in the next section is slightly better than the pivotal bootstrap on the right, but is vastly inferior on the left. Indeed, on the left the saddlepoint method is even worse than an approximation based on the assumption that the sample $\mathscr{X}$ is selected from a bivariate normal distribution; specifically,

$$\Pr\left\{(\bar{X}^{(1)} - \theta_1)\Big/\left[(1 - \hat{\rho}^2) n \hat{\sigma}_1^2/(n-2)\right]^{1/2} \leq y\right\} \simeq \Pr\{T_{n-2} \leq y\},$$

where T_{n-2} has a t-distribution with $n - 2$ degrees of freedom.

Our main conclusion from this example is that pivoting is important and may in practice lead to a substantial improvement in performance. This conclusion is based on the suggestion in Section 4 that the saddlepoint approach accurately approximates the nonpivotal bootstrap with a suitable choice of bandwidth. As in other, more traditional applications of the bootstrap, the advantages of pivoting are more substantial for more highly skewed distributions. The exponential population in this example has relatively large skewness.

4. Saddlepoint methods.

In Section 5.1 we note that the problem of bandwidth selection is more complicated for the nonpivotal estimate, $\hat{q}^{\dagger}$, because the relevant limiting conditional distribution depends on an unknown parameter, ρ. Recall that when $K(u) = I\{|u| < 1\}/2$, $\hat{q}^{\dagger}$ is an approximation to $\tilde{q}$ in (2.2) based on a finite number of resamples. In the setting described in Section 3, the nonpivotal estimate $\tilde{q}$ may be approximated directly (i.e., without resampling) using the saddlepoint method advocated by Davison and Hinkley (1988), Section 5. This method may be described as follows.

Let $a(\bar{X}^*)$ denote the density of the discrete bootstrap distribution of the bivariate sample mean $\hat{\theta} = (\bar{X}^{(1)}, \bar{X}^{(2)})^T$, and let $f(\bar{X}^*; \beta)$ denote the "tilted" bootstrap distribution

$$f(\bar{X}^*; \beta) = a(\bar{X}^*) e^{n\{\beta^T \bar{X}^* - \kappa(\beta)\}}, \tag{4.1}$$

where $\beta = (\beta_1, \beta_2)^T$ and $\kappa(\beta) = \log\{(1/n)\Sigma_{i=1}^{n} e^{\beta^T X_i}\}$. Then an approximation to (2.2) is obtained by applying Skovgaard's (1987) saddlepoint formula for the conditional distribution of $\bar{X}^{(1)}$ given $\bar{X}^{(2)} = y$ when $\beta_1 = 0$ under the model (4.1). Specifically, this yields

$$\tilde{q} \approx \Phi(\zeta) + \phi(\zeta)\left\{\frac{1}{\zeta} - \frac{1}{\xi}\right\}, \tag{4.2}$$

where $\phi = \Phi'$ and Φ is the standard normal distribution function;

$$\zeta = \operatorname{sgn}(\hat{\beta}_1)\Big[-2n\Big\{\hat{\beta}_{2(0)}, y - \kappa\big(\hat{\beta}_{(0)}\big) - \hat{\beta}_1 x - \hat{\beta}_2 y + \kappa\big(\hat{\beta}\big)\Big\}\Big]^{1/2};$$

$\xi = \hat{\beta}_1\{|j(\hat{\beta})|/|j_{22}(\hat{\beta}_{(0)})|\}^{1/2}$; and $j(\beta) = \partial^2\kappa/\partial\beta\,\partial\beta^T$, $j_{22}(\beta) = \partial^2\kappa/\partial\beta_2^2$, $\hat{\beta}$ solves $\partial\kappa/\partial\beta^T|_{\beta=\hat{\beta}} = (x, y)$ and $\hat{\beta}_{(0)}^T = (0, \hat{\beta}_{2(0)})$ solves $\partial\kappa/\partial\beta_2|_{\beta=\hat{\beta}_{(0)}} = y$. A modified form of the approximation is needed if $\hat{\beta}_1 = 0$. In their example Davison and Hinkley found the saddlepoint approximation (4.2) worked well for a bandwidth $h = 0.25\hat{\sigma}_2$.

Table 3 shows the accuracy of the saddlepoint approximation to $\tilde{q}$ for the first sample of size $n = 10$ in the simulation study of Section 3. In Table 3 the "exact" values of $\tilde{q}$ are given corresponding to quantiles obtained by inverting the saddlepoint conditional distribution for $\hat{\theta}_1^*$ given $\hat{\theta}_2^* = 1$ with $h = 0.05$, 0.10 and 0.25 times $\hat{\sigma}_2$, the standard error of $\hat{\theta}_2 = \bar{X}^{(2)}$. By "exact" we mean that the values were the fraction of times that $\hat{\theta}_1^*$ fell below the saddlepoint quantile out of those resamples for which $1 - h \le \hat{\theta}_2^* \le 1 + h$ in a total of 10^6 uniform resamples. Table 3 suggests that the saddlepoint approximation to $\tilde{q}$ is good for $h = 0.05$ and 0.10 times $\hat{\sigma}_2$, but not for $h = 0.25$ $\hat{\sigma}_2$, the value suggested by Davison and Hinkley. Thus, our numerical results corroborate those of Davison and Hinkley (1988), although a different bandwidth was required. It seems we are left with little insight into the smoothing mechanism implicit in the saddlepoint method.

TABLE 3
Comparison of the saddlepoint approximation (4.3) *with the nonpivotal approximation* (2.2) *at various bandwidths*

Saddlepoint percentage points	**Percentage points of $\tilde{q}$**		
	$h = 0.05\hat{\sigma}_2$	$h = 0.10\hat{\sigma}_2$	$h = 0.25\hat{\sigma}_2$
0.005	0.00522	0.00534	0.00943
0.010	0.00809	0.00925	0.01850
0.025	0.02347	0.02371	0.03897
0.050	0.06077	0.05563	0.06943
0.100	0.11007	0.10396	0.12112
0.200	0.21388	0.20023	0.22487
0.800	0.80490	0.80081	0.78312
0.900	0.89463	0.89148	0.87970
0.950	0.94940	0.94359	0.93155
0.975	0.97131	0.97316	0.96351
0.990	0.98905	0.98945	0.98397
0.995	0.99452	0.99401	0.98979

5. Large-sample theory.

5.1. *Main results and summary.* Recall from Section 2.4 that $p = p(u, w) = \Pr(U \le u | W = w)$, where

$$U = (1 - \hat{\rho}^2)^{-1/2}(V - \hat{\rho}w), \qquad V = (\hat{\theta}_1 - \theta_1)/\hat{\sigma}_1, \qquad W = (\hat{\theta}_2 - \theta_2)/\hat{\sigma}_2$$

and $\hat{\rho}$ is a sample estimate of the asymptotic coefficient of correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$. We begin this section by outlining properties of the bootstrap approximant $\hat{p}^{\dagger}$, defined in Section 2.4.

As noted in Section 2.4, the definition of U is deliberately chosen so that, as $n \to \infty$, $p \to \Phi(u)$, where Φ is the standard normal distribution function. In more detail, p admits an Edgeworth expansion of the form

$$(5.1) \qquad p(u, w) = \Phi(u) + \sum_{l=1}^{m} n^{-l/2} Q_l(u|w)\phi(u) + O(n^{-(m+1)/2}),$$

where $\phi = \Phi'$ and $Q_l(u|w)$ is a polynomial of degree $3l - 1$ in u, with coefficients depending on population moments. Section 5.2 will describe the origins of this expansion. Replacing the population moments in Q_l by sample moments, giving a new polynomial $\hat{Q}_l$, we obtain from (5.1) an Edgeworth approximation to p:

$$(5.2) \qquad \check{p}_m(u, w) = \Phi(u) + \sum_{l=1}^{m} n^{-l/2} \hat{Q}_l(u|w)\phi(u).$$

At this stage in our argument (5.2) is no more than a definition of the empirical Edgeworth approximation $\check{p}_m$. Section 5.3 will prove that, apart from random fluctuations arising from the resampling procedure, and excepting terms of order $n^{-(m+1)/2} + h^2$, the bootstrap approximant $\hat{p}^{\dagger}$ is identical to $\check{p}_m$. In this sense, the bootstrap estimate can be viewed as an empirical Edgeworth approximation, much as in more standard problems of bootstrap inference.

Recall from Section 2.4 that $\hat{p}^\dagger = \hat{\lambda}_1^\dagger/\hat{\lambda}_2^\dagger$, where $\hat{\lambda}_1^\dagger$ and $\hat{\lambda}_2^\dagger$ are given by (2.5) and (2.6) and are computed using kernel methods. Let us take the kernel K to be of second order, which is the most common type in practice. Then $\int K = 1$, $\int yK(y)\,dy = 0$ and $2\kappa_1 = \int y^2K(y)\,dy \neq 0$; for example, K could be a density function such as the standard normal density. In Section 5.3 we shall prove that for each $m \geq 1$,

$$\begin{aligned}\hat{p} &= E'\big(\hat{\lambda}_1^\dagger\big)/E'\big(\hat{\lambda}_2^\dagger\big) = \hat{\lambda}_1/\hat{\lambda}_2\\ &= \check{p}_m + h^2\kappa_1\psi(u,w) + O_p\big(h^3 + n^{-1/2}h^2 + n^{-(m+1)/2}\big),\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}\hat{\lambda}_1 &= E'\big(\hat{\lambda}_1^\dagger\big) = h^{-1}E'\big[I(U^* \leq u)K\{(w - W^*)/h\}\big],\\ \hat{\lambda}_2 &= E'\big(\hat{\lambda}_2^\dagger\big) = h^{-1}E'\big[K\{(w - W^*)/h\}\big],\end{aligned} \tag{5.4}$$

$$\psi(u,v) = \Big\{2\rho\big(1 - \rho^2\big)^{-1/2}w - \rho^2\big(1 - \rho^2\big)^{-1}u\Big\}\phi(u) \tag{5.5}$$

and ϕ denotes the standard normal density.

These results imply that the error in the bootstrap approximation $\hat{p}^\dagger(u,w)$ to $p(u,w)$ can be as little as $O_p(n^{-1})$, for judicious choice of B and h. That represents a significant improvement on the error in the normal approximation, $p(u,w) \simeq \Phi(u)$, which is of size $n^{-1/2}$. To check our claim about the order of $\hat{p}^\dagger - p$, note that since $\hat{Q}_l - Q_l = O_p(n^{-1/2})$, then, subtracting (5.1) and (5.2), $p - \check{p}_m = O_p(n^{-1})$ for $m \geq 1$. Hence by (5.3), $\hat{p} - p = O_p(n^{-1} + h^2)$. As noted in the Appendix, the Monte Carlo approximation $\hat{p}^\dagger$ of $\hat{p}$ is subject to additional random fluctuations of order $(Bh)^{-1/2}$, arising from the Monte Carlo resampling. Therefore,

$$\hat{p}^\dagger - p = O_p\big\{n^{-1} + h^2 + (Bh)^{-1/2}\big\}.$$

Thus, provided B and h are chosen so that $h^4 + (Bh)^{-1} = O(n^{-1})$, we have $\hat{p}^\dagger - p = O_p(n^{-1})$.

This result, that the error in the bootstrap approximation can be as small as $O_p(n^{-1})$, is not available for the nonpivotal methods introduced in Section 2.3. We shall demonstrate why by discussing the problem of estimating $q = \Pr(\hat{\theta}_1 \leq x|\hat{\theta}_2 = y)$.

Let σ_k^2 denote the asymptotic variance of $\hat{\theta}_k$ and define $s = (x - \theta_1)/\sigma_1$, $t = (y - \theta_2)/\sigma_2$, $S = (\hat{\theta}_1 - \theta_1)/\sigma_1$, $T = (\hat{\theta}_2 - \theta_2)/\sigma_2$. Then S, conditional on $T = t$, is asymptotically normal $N(\rho t, 1 - \rho^2)$. Therefore,

$$q = \Pr(S \leq s|T = t) = \Phi\Big\{\big(1 - \rho^2\big)^{-1/2}(s - \rho t)\Big\} + \cdots,$$

where the quantity represented by "..." comprises terms of order $n^{-1/2}$ and smaller in an Edgeworth expansion. This result plays the role of (5.1) in the present context, the crucial difference being that here, the first term depends on the unknown ρ. In the version of (5.2) for this case, the first term is changed to $\Phi\{(1-\hat{\rho}^2)^{-1/2}(s-\hat{\rho}t)\}$. Thus, when the versions of (5.1) and (5.2) are subtracted, we obtain

$$\Phi\Big\{\big(1-\rho^2\big)^{-1/2}(s-\rho t)\Big\} - \Phi\Big\{\big(1-\hat{\rho}^2\big)^{-1/2}(s-\hat{\rho}t)\Big\} = O_p(n^{-1/2}),$$

rather than simply $O_p(n^{-1})$ as was formerly the case. This quantity is of size $n^{-1/2}$, not n^{-1}, since $\hat{p}-p$ is of size $n^{-1/2}$.

5.2. *Expansion of p.* Under mild assumptions, for example Cramér's condition [e.g., Bhattacharya and Rao (1976), page 207], a moment condition on the parent population and the assumption introduced in Section 2.5 that $\theta_k = g_k(\mu)$ for a smooth function g_k, the joint distribution of $U = (1-\hat{\rho}^2)^{-1/2}\{(\hat{\theta}_1-\theta_1)\hat{\sigma}_1^{-1} - \hat{\rho}w\}$ and $W = (\hat{\theta}_2-\theta_2)/\hat{\sigma}_2$ admits an Edgeworth expansion of arbitrarily high order:

$$\begin{aligned}\sup_{C\in\mathscr{C}}\big(1+\|\partial C\|^2\big)\bigg|&\Pr\{(U,W)\in C\}\\ &-\int_C\bigg\{\phi_{\rho,w}(\xi,\eta)+\sum_{l=1}^{m}n^{-l/2}Q_l(\xi,\eta)\phi_{\rho,w}(\xi,\eta)\bigg\}\,d\xi\,d\eta\bigg|\\ &= O\big(n^{-(m+1)/2}\big).\end{aligned}\tag{5.6}$$

Here, $\mathscr{C}$ denotes the class of all convex sets C, $\|\partial C\|$ is the Euclidean distance of the boundary of C from the origin [Bhattacharya and Rao (1976), page 170], $\phi_{\rho,w}$ is the bivariate $N\{-(1-\rho^2)^{-1/2}\rho w, 0; (1-\rho^2)^{-1}, 1; \rho\}$ density, and Q_l is a polynomial of degree $3l$ whose coefficients depend on population moments. If, in addition, the conditional distribution of U given W is well defined, for which we ask that the parent population be continuous, then an expansion of $p = \Pr(U \le u|W=w)$ is obtainable directly from the Edgeworth expansion at (5.6):

$$\begin{aligned}p &= \Pr(U \le u|W=w)\\ &= \frac{\int_{-\infty}^{u}\big\{\phi_{\rho,w}(\xi,w)+\sum_{l=1}^{m}n^{-l/2}Q_l(\xi,w)\phi_{\rho,w}(\xi,w)\big\}\,d\xi}{\int_{-\infty}^{\infty}\big\{\phi_{\rho,w}(\xi,w)+\sum_{l=1}^{m}n^{-l/2}Q_l(\xi,w)\phi_{\rho,w}(\xi,w)\big\}\,d\xi} + O\big(n^{-(m+1)/2}\big)\\ &= \Phi(u)+\sum_{l=1}^{m}n^{-l/2}Q_l(u|w)\phi(u)+O\big(n^{-(m+1)/2}\big),\end{aligned}$$

say. Here, Φ and ϕ are the univariate standard normal distribution and density functions, respectively, and $Q_l(u|w)$ is a polynomial of degree $3l-1$ in u. This establishes (5.1).

5.3. *Expansions of $\hat{p}$.* Under the conditions leading to (5.6), that is, Cramér's condition, a moment condition and the "smooth function model" for the unknowns θ_k, the bootstrap version of (5.6) is valid:

$$\sup_{C \in \mathscr{C}} (1 + \|\partial C\|^2) \Bigg| \Pr'\{(U^*, W^*) \in C\} \\ - \int_C \left\{ \phi_{\hat{\rho}, w}(\xi, \eta) + \sum_{l=1}^{m} n^{-l/2} \hat{Q}_l(\xi, \eta) \phi_{\hat{\rho}, w}(\xi, \eta) \right\} d\xi\, d\eta \Bigg| \\ = O_p(n^{-(m+1)/2}). \tag{5.7}$$

Here, $\hat{\rho}$ is a sample estimate of ρ, $\hat{\sigma}_k^*$ and $\hat{\rho}^*$ denote the versions of $\hat{\sigma}_k$ and $\hat{\rho}$ computed for a resample $\mathscr{X}^*$ rather than the sample $\mathscr{X}$, $U^* = (1 - \hat{\rho}^{*2})^{-1/2}\{(\hat{\theta}_1^* - \hat{\theta}_1)\hat{\sigma}_1^{*-1} - \hat{\rho}^* w\}$, $W^* \doteq (\hat{\theta}_2^* - \hat{\theta}_2)/\hat{\sigma}_2^*$, and $\hat{Q}_l$ is obtained from Q_l on replacing population moments by corresponding sample moments.

Define $\hat{f}(\xi, \eta) = \phi_{\hat{\rho}, w}(\xi, \eta) + \sum_{1 \le l \le m} n^{-l/2} \hat{Q}_l(\xi, \eta) \phi_{\hat{\rho}, w}(\xi, \eta)$, which is an Edgeworth approximation to the density f of (U, W). Put $\hat{f}_W(\eta) = \int \hat{f}(\xi, \eta)\, d\xi$. We may deduce from (5.7) that

$$\hat{\lambda}_1 = h^{-1} \int_{\xi \le u,\, -\infty < \eta < \infty} K\{(w - \eta)/h\} \hat{f}(\xi, \eta)\, d\xi\, d\eta + O_p(h^{-1} n^{-(m+1)/2}),$$

$$\hat{\lambda}_2 = h^{-1} \int_{-\infty < \eta < \infty} K\{(w - \eta)/h\} \hat{f}_W(\eta)\, d\eta + O_p(h^{-1} n^{-(m+1)/2}).$$

Changing variable from η to $\eta' = (w - \eta)/h$ in each integral and then Taylor expanding the integrands, we deduce that

$$\hat{\lambda}_1 = \int_{\xi \le u} \hat{f}(\xi, w)\, d\xi + h^2 \kappa_1 \int_{\xi \le u} \hat{f}^{\langle 0, 2 \rangle}(\xi, w)\, d\xi + O_p(h^3 + h^{-1} n^{-(m+1)/2}),$$

$$\hat{\lambda}_2 = \hat{f}_w(w) + h^2 \kappa_1 \hat{f}_2^{\langle 2 \rangle}(w) + O_p(h^3 + h^{-1} n^{-(m+1)/2}).$$

Here, $a^{\langle 2 \rangle}$ denotes the second derivative of a univariate function a, and $a^{\langle 0, 2 \rangle}$ denotes the second derivative with respect to the second variable in a bivariate function a. Noting that $\hat{f} - \phi_{\rho, w}$ and $\hat{f}_W - \phi$ are both of order $n^{-1/2}$, we obtain

$$\hat{\lambda}_1 / \hat{\lambda}_2 = \check{p}_{(m)} + h^2 \kappa_1 \left\{ \phi(w)^{-1} \int_{-\infty}^{u} \phi_{\rho, w}^{\langle 0, 2 \rangle}(\xi, w)\, d\xi \right. \\ \left. - \phi(w)^{-2} \phi^{\langle 2 \rangle}(w) \int_{-\infty}^{u} \phi_{\rho, w}(\xi, w)\, d\xi \right\} \\ + O_p(h^3 + n^{-1/2} h^2 + h^{-1} n^{-(m+1)/2}), \tag{5.8}$$

where $\check{p}_{(m)} = \hat{f}_W(w)^{-1} \int_{\xi \le u} \hat{f}(\xi, w)\, d\xi$.

Let $m' \ge 1$ be given. Since h decreases to 0 no faster than n^{-c} for some $c > 0$, then we may choose $m > m'$ so large that $h^{-1} n^{-(m+1)/2} = O(n^{-(m'+1)})$. Furthermore, if $\check{p}_m$ is the Edgeworth approximation defined at (5.2), then $\check{p}_{(m)} - \check{p}_m = O(n^{-(m+1)/2})$. Hence result (5.8) implies (5.3) with m in the latter replaced by m'.

APPENDIX

We treat only estimation of $p = \Pr(U \le u \mid W = w)$; other cases are similar. Let $U^\dagger, W^\dagger, M_j^\dagger$ denote generic versions of $U_i^\dagger, W_i^\dagger, M_{ij}^\dagger$, respectively, and write U^*, W^*, M_j^* for versions of the former variables under uniform resampling rather than importance resampling. Define $\hat\lambda_1^\dagger$ and $\hat\lambda_2^\dagger$ as at (2.5) and (2.6), let $\hat\lambda_1$ and $\hat\lambda_2$ be as at (5.4) and (5.5), and put $\eta_j = -\log(n\pi_j)$,

$$\Delta_1 = \frac{1}{Bh}\sum_{i=1}^{B} I(U_i^\dagger \le u) K\left\{\frac{w - W_i^\dagger}{h}\right\}\exp\left(\sum_{j=1}^{n} M_{ij}^\dagger \eta_j\right) - \hat\lambda_1,$$

$$\Delta_2 = \frac{1}{Bh}\sum_{i=1}^{B} K\left\{\frac{w - W_i^\dagger}{h}\right\}\exp\left(\sum_{j=1}^{n} M_{ij}^\dagger \eta_j\right) - \hat\lambda_2.$$

Then $E'(\Delta_k) = 0$ for $k = 1, 2$ and

$$\hat p^\dagger = \left(\hat\lambda_1 + \Delta_1\right)/\left(\hat\lambda_2 + \Delta_2\right) = \hat p + \hat\lambda_1^{-1}(\Delta_1 - \hat p\Delta_2) + o_p\{(Bh)^{-1/2}\},$$

where $\hat p = \hat\lambda_1/\hat\lambda_2$. The conditional variance γ of $(Bh)^{1/2}(\Delta_1 - \hat p\Delta_2)$ is asymptotic to

$$\begin{aligned}
& h^{-1}E'\left[\{I(U^\dagger \le u) - \hat p\}^2 K\{(w - W^\dagger)/h\}^2 \exp\left(2\sum_{j=1}^{n} M_j^\dagger \eta_j\right)\right] \\
& \quad = h^{-1}E'\left[\{I(U^* \le u\} - \hat p\}^2 K\{(w - W^*)/h\}^2 \exp\left(\sum_{j=1}^{n} M_j^* \eta_j\right)\right] \\
\text{(A.1)} \quad & \quad \sim \left(\int K^2\right) f_W(w) E'\left[\{I(U^* \le u) - \hat p\}^2\right. \\
& \qquad\qquad \left. \times \exp\left(\sum_{j=1}^{n} M_j^* \eta_j\right) \middle| w - h \le W^* \le w + h\right].
\end{aligned}$$

Now, $U^* \simeq U_0^* = (1 - \rho^2)^{-1/2}\sum M_j^*\varepsilon_{1j} - (1 - \rho^2)^{-1/2}\rho w$, $W^* \simeq W_0^* = \sum M_j^*\varepsilon_{2j}$. Define $T_0^* = \sum M_j^*\eta_j$ and $\eta_j = A_1\varepsilon_{1j} + A_2\varepsilon_{2j} + A_3\xi_j + C$, where $(\xi_1, \dots, \xi_n)$ is orthogonal to each of $(1, \dots, 1)$, $(\varepsilon_{11}, \dots, \varepsilon_{1n})$ and $(\varepsilon_{21}, \dots, \varepsilon_{2n})$, $\sum \xi_j^2 = 1$, and C is chosen to ensure that $\sum \pi_j = 1$. Then $C = (2n)^{-1}s^2 + o(n^{-1})$, where $s^2 = A_1^2 + A_2^2 + A_3^2 + 2A_1A_2\rho$, and $E'(U_0^*) = -(1 - \rho^2)^{-1/2}\rho w$, $E'(W_0^*) = 0$, $E'(T_0^*) = nC$, $\mathrm{Var}'(U_0^*) = (1 - \rho^2)^{-1/2} + o(1)$, $\mathrm{Var}'(W_0^*) = 1 + o(1)$, $\mathrm{Var}'(T_0^*) = s^2 + o(1)$, $\mathrm{Cov}'(U_0^*, W_0^*) = (1 - \rho^2)^{-1/2}\rho + o(1)$, $\mathrm{Cov}'(T_0^*, U_0^*) = (1 - \rho^2)^{-1/2}(A_1 + A_2\rho) + o(1)$, $\mathrm{Cov}'(T_0^*, W_0^*) = (1 - \rho^2)^{-1/2}(A_1\rho + A_2) + o(1)$.

Let (T_0, U_0, W_0) be normally distributed with mean $(s^2/2, -(1 - \rho^2)^{-1/2}\rho w, 0)$ and variance matrix

$$\begin{bmatrix}
s^2 & (2 - \rho^2)^{-1/2}(A_1 + A_2\rho) & A_1\rho + A_2 \\
* & (1 - \rho^2)^{-1} & (1 - \rho^2)^{-1/2}\rho \\
* & * & 1
\end{bmatrix}.$$

Then by (A.1)

$$\gamma\Big/\Big\{\Big(\int K^2\Big) f_W(w)\Big\} \sim E\Big[\{I(U_0 \le u) - p\}^2 e^{T_0}\big|W_0 = w\Big] = \beta,$$

where

$$\beta = (1 - 2p)E\{I(U_0 \le u)e^{T_0}|W_0 = w\} + p^2E(e^{T_0}|W_0 = w).$$

Conditional on $W_0 = w$, (T_0, U_0) is normally distributed with mean $(s^2/2 + (A_1\rho + A_2)w, 0)$, variances $(A_3^2 + A_1^2(1 - \rho^2), 1)$ and covariance $A_1(1 - \rho^2)^{1/2}$. Therefore,

$$\beta = \Big[(1 - 2p)\Phi\Big\{u - A_1(1 - \rho^2)^{1/2}\Big\} + p^2\Big]$$
$$\times \exp\Big\{(A_1\rho + A_2)w + A_3^2 + A_1^2(1 - \rho^2) + \tfrac{1}{2}(A_1\rho + A_2)^2\Big\}.$$

It follows that β is minimized by taking $A_3 = 0$, and also $A_2 = -A_2\rho$ if $w = 0$. This gives the formulae at (2.9) and (2.10), when it is noted that $p - \Phi(u) \to 0$ as $n \to \infty$.

If we stipulate that each $\pi_i = n^{-1}$, then $A_1 = A_2 = A_3 = 0$, whence $\beta = (1 - 2p)\Phi(u) + p^2 \to \Phi(u)\{1 - \Phi(u)\}$.

REFERENCES

BHATTACHARYA, R. N. AND RAO, C. R. (1976). *Normal Approximation and Asymptotic Expansions.* Wiley, New York.

DAVISON, A. C. (1988). Discussion of "Bootstrap methods" by D. V. Hinkley and "A review of bootstrap confidence intervals" by T. J. DiCiccio and J. P. Romano. *J. Roy. Statist. Soc. Ser. B* **50** 356–357.

DAVISON, A. C. and HINKLEY, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75** 417–431.

DICICCIO, T. J. and ROMANO, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 338–370.

HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16** 927–953.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HINKLEY, D. V. (1988). Bootstrap methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 321–370.

JOHNS, M. V., JR. (1988). Importance resampling for bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **83** 709–714.

KENDALL, M. G. and STUART, A. (1979). *The Advanced Theory of Statistics* **2**. Griffin, London.

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

SKOVGAARD, I. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24** 275–287.

CENTER FOR MATHEMATICS AND
ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA A.C.T. 2601
AUSTRALIA