Bootstrap Confidence Intervals

by

G. J. Babu
Pennsylvania State University

and

A. Bose
Purdue University

Technical Report #87-54

Bootstrap Confidence Intervals

by

G. J. Babu
Pennsylvania State University

and

A. Bose
Purdue University

Abstract:

Nonparametric confidence bounds are obtained for a wide class of statistics using bootstrap. These results improve the errors in the probability estimates of the confidence intervals over the ones obtained by the normal approximation theory unconditionally.

Key Words and Phrases: Normal approximation, nonparametric confidence intervals, Cramer's condition, Edgeworth expansion, bootstrap.

AMS 1985 Subject Classification: Primary 62G15, Secondary 62F99.

◇

1. Introduction:

Let θ be a real valued parameter for which a confidence interval is desired based on i.i.d. observations $X_1, X_2, \ldots, X_n$ from the unknown population distribution function F on $\mathbb{R}^k$, $k \geq 1$. An approximate solution is obtained as follows. Let $\theta_n = \theta_n(X_1, \ldots, X_n)$ be an estimator of θ. Consider $T_n = T_n(X_1, \ldots, X_n) = (\theta_n - \theta)/s_n$ where s_n^2 is the estimated variance of θ_n. In a wide variety of situations, the distribution function H_n of T_n approaches a standard normal distribution as $n \to \infty$. Hence an approximate $100(1 - \alpha)\%$ confidence interval for θ is given by $\theta_n \pm z_{\alpha/2} s_n$ where $z_{\alpha/2}$ is the upper $100(1 - \alpha/2)\%$ point of the standard normal distribution. The accuracy of this confidence interval depends on how good the normal approximation is. But in most cases it is not better than $0(n^{-1/2})$ (recall the Berry Esseen bound). In certain very simple cases (e.g. normal population with known variances) the accuracy might be better.

The bootstrap technique replaces the normal by a certain data dependent distribution as the approximating distribution. This method was introduced by Efron (1979). To describe this method, let $T_n = T = T(X, F)$ be a statistic based on a sample $X = (X_1, \ldots, X_n)$ from F. Let $X^* = (X_1^*, \ldots, X_n^*)$ be a simple random sample with replacement from $(X_1, \ldots X_n)$, i.e. X_i^* are i.i.d. from F_n where F_n is the e.d.f. of $X_1, \ldots, X_n$. Let $T_n^* = T(X^*, F_n)$ and let H_n^* be the distribution function of T_n^*. Note that given the data $X_1, \ldots, X_n$, H_n^* can be explicitly computed or can be approximated to any desired degree of accuracy by drawing repeated sets of observations $(X_{i1}^*, \ldots, X_{in}^*)$, $i = 1, 2, \ldots$ from F_n. The bootstrap idea is to approximate H_n by H_n^*.

For a wide class of statistics T_n and a wide class of distribution functions F, this approximation has a great degree of accuracy. It essentially corrects for the skewness of the sampling distribution. See Bickel and Freedman (1981), Singh (1981), Babu and Singh (1983, 1984a,b) and Babu (1984). The essence of these results is that under proper smoothness conditions on F and T, $n^{1/2}\|H_n - H_n^*\| = \sup_x n^{1/2}|H_n(x) - H_n^*(x)| \to 0$ a.s. as $n \to \infty$. The a.s. here means that the above result holds for almost every sequence $X_1, X_2, \ldots$. The $100(1 - \alpha)\%$ confidence interval for θ based on this approximation is $(\theta_n - h_n(\alpha/2)s_n, \theta_n + h_n(1 - \alpha/2)s_n)$ where $h_n(\alpha/2)(h_n(1 - \alpha/2))$ is the $100\alpha/2\%(100(1 - \alpha/2)\%)$ point of H_n^* and the accuracy of this is now $o(n^{-1/2})$. One sided confidence intervals can be built in a similar way. This is already an improvement over the normal approximation. However note that a different type of error has appeared (the existence of "almost sure").

There are ways to improve the normal approximation directly. This depends on higher order expansions (Edgeworth expansions) of H_n. Suppose that H_n has an expansion of the form

$$H_n(x) = \Phi(x) + n^{-1/2} \int_{-\infty}^{x} p(y)\phi(y)dy + 0(n^{-1})$$

where Φ and ϕ denote the distribution and density functions of a standard normal variable. Then the modified statistics $\tilde{T}_n = T_n + n^{-1/2}p(T_n, F)$ has an expansion for its distribution function $\tilde{H}_n$ and $\tilde{H}_n(x) = \Phi(x) + 0(n^{-1})$. Then one can hope to invert the tails of $\tilde{T}_n$ to get a confidence interval for θ — the quantities unknown in p (usually p is a polynomial with

coefficients as functions of the moments) have to be replaced by their estimates. With some more conditions, the distribution of $\tilde{T}_n$ can be bootstrapped and that has accuracy $o(n^{-1})$. For details see Abramovitch and Singh (1985). So in general the bootstrap outperforms the classical normal approximation. However the disadvantage of the above procedure is that p has to be known.

In the special case of multivariate normal populations, Efron (1985) used bootstrap techniques to obtain two sided confidence intervals. He terms it the bias corrected percentile method. But this method does not seem to perform well for obtaining one sided intervals. An improved method called BC_a method was introduced in Efron (1987). It gives intervals which are second order correct and work under existence of certain transformations. In the next section we will show that in the i.i.d. situation where Edgeworth expansion exists for H_n the bootstrap technique can be used to get confidence intervals of the accuracy $0(n^{-1}(\log n)^\alpha)$. The advantages of this procedure are (i) it does not depend explicitly on the form of p but uses only its existence (which can be guaranteed under fairly general conditions), (ii) we do away with "almost sure" conditions and (iii) one sided confidence intervals pose no problems.

2. The Main Results:

We will use the following notations. For any distribution function G and any $0 < \alpha \leq 1$, $G^{-1}(\alpha) = \inf(x: G(x) \geq \alpha)$, $G^{-1}(0) = \lim_{\alpha \downarrow 0} G^{-1}(\alpha)$ and $j(G) = \sup_x (G(x) - G(x-))$, the maximum jump of the distribution function G. ϕ_V will denote the density function of a normal variable with mean 0 and dispersion matrix V.

Our first lemma shows how an estimate of the difference between H_n and H_n^* can be transferred into an *unconditional* probability statement about T. This lemma is the main tool of the paper.

<u>Lemma 2.1</u>: Suppose that for a positive sequence of numbers (ϵ_n),

$$P(\|H_n - H_n^*\| \geq \epsilon_n) \leq \epsilon_n.$$

Then $\sup_{0 \leq \alpha \leq 1} |P(T < H_n^{*^{-1}}(\alpha)) - \alpha| \leq 2\epsilon_n + j(H_n)$

<u>Proof</u>: Let $A_n = \{|H_n(T) - H_n^*(T)| \geq \epsilon_n\}$. Then

$$
\begin{aligned}
P(T < H_n^{*^{-1}}(\alpha)) - \alpha &= P(H_n^*(T) < \alpha) - \alpha \\
&= P(H_n(T) < \alpha + (H_n(T) - H_n^*(T))) - \alpha \\
&\leq P(A_n) + P(H_n(T) < \alpha + \epsilon_n) - \alpha \\
&\leq P(A_n) + \epsilon_n + j(H_n) \\
&\leq 2\epsilon_n + j(H_n).
\end{aligned}
$$

Similarly $\alpha - P(T < H_n^{*^{-1}}(\alpha)) \leq 2\epsilon_n + j(H_n)$. This proves the lemma.

The most commonly used statistics are generally of the type

$$T = n^{1/2}(K(\overline{Z}) - K(\mu))/\nu(n^{-1} \sum_{i=1}^{n} \lambda(Z_i)) \text{ where}$$

$$\overline{Z} = n^{-1} \sum_{i=1}^{n} Z_i,$$

Z_i's are i.i.d. $\mathbb{R}^k$ valued random variables. K is a real valued function on $\mathbb{R}^k$, λ is a function from $\mathbb{R}^k$ to $\mathbb{R}^g$ and ν is a real valued function on $\mathbb{R}^g$. An example is the t statistic,

$$T = n^{1/2}(\overline{X}_n - \mu)s_n^{-1}$$

where X_i's are i.i.d. real valued random variables.

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i, \quad s_n^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

In this case take

$$Z_i = X_i$$
$$R(x) = x$$
$$\lambda(x) = (x^2, x)$$
$$\nu(x, y) = \max(0, (x - y^2)^{1/2})$$

Our main aim is to be obtain results for such type of statistics. The first step will be to prove the result for linear functions of $\overline{Z}$. The general case will involve expansions of the function K. The accuracy will depend on the smoothness of K. To treat the general case we need the following lemma which is an extension of Lemma 3 of Babu and Singh (1984a).

<u>Lemma 2.2</u>: Let $l = (l_1, \ldots, l_k)$ be a vector, $L = (L_{ij})$ be a $k \times k$ matrix and $A = (a_{ijs})$ be a $k \times k \times k$ 3 dimensional array. Let Q_1 and Q_2 be two polynomials in k variables with coefficients (a_ν). Let V be a $k \times k$ positive definite matrix and $(u_{ij}) = V^{-1}$. Let $M \geq \max\{|l_i|, |L_{ij}|, |a_{ijs}|, |v_{ij}|, |u_{ij}|, |a_\nu|\}$. Let $|l_k| > l_0 > 0$ and $b_n = (l_k n^{1/2})^{-1}$. Then there exists polynomials p_i in one variable whose coefficients are continuous functions of $l_i, L_{ij}, a_{ijs}, u_{ij}, v_{ij}$ and a_ν such that

(i)
$$\int_{\{z:\ l'z + n^{-1/2} z' Lz < u(l'Vl)^{1/2}\}} (1 + n^{-1/2} Q_1(z)) \phi_V(z) dz$$
$$= \int_{-\infty}^{u} (1 + b_n p_1(y)) \phi(y) dy + 0(n^{-1}(\log n)^{3/2})$$

3

$$\int_{\{z:\ l'z+n^{-1/2}z'Lz+n^{-1}\sum_{i,j,s}z_iz_jz_sa_{ijs}<u(l'Vl)^{1/2}\}}(1+n^{-1/2}Q_1(z)+n^{-1}Q_2(z))\phi_V(z)dz$$

$$=\int_{-\infty}^{u}(1+b_np_1(y)+b_n^2p_2(y))\phi(y)dy+0(n^{-3/2}(\log\ n)^\beta)$$

and $0(\cdot)$ terms depend only on M and l_0, and β depends on the degrees of Q_1 and Q_2.

Proof: The proof of (i) is contained in the proof of Lemma 3 of Babu and Singh (1984a). The reader can also get an idea of the details from the following proof of (ii). So we will not prove (i) here.

Define

$$v = u(l'Vl)^{1/2}$$
$$h(z) = l'z$$
$$g(z) = l'z + n^{-1/2}z'Lz + n^{-1}\sum_{i,j,s}z_iz_jz_sa_{ijs}$$
$$z = (\tilde{z}, z_k)$$
$$y = z_k + b_nz'Lz + n^{-1/2}b_n\sum_{i,j,s}z_iz_jz_sa_{ijs}$$
$$r(z) = (\tilde{z}, y)$$

Then clearly $h(r(z)) = g(z)$.

Let $C_n = \{z:\ \|z\| < C(\log n)^{1/2}\}$ where C is a large constant. On C_n, (q_i's and $\tilde{q}_i$'s are polynomials),

$$y - z_k = b_nz'Lz + n^{-1/2}b_n\sum_{i,j,s}z_iz_jz_sa_{ijs}$$
$$= b_n\{q_1(\tilde{z}) + n^{-1}\tilde{q}_1(\tilde{z}) + z_k(q_2(\tilde{z}) + n^{-1}\tilde{q}_2(\tilde{z}))$$
$$+ z_k^2(q_3(\tilde{z}) + n^{-1}\tilde{q}_3(\tilde{z}))\}$$
$$= b_nq_4(\tilde{z}, y) + b_n^2\tilde{q}_4(\tilde{z}, y) + 0((\log n)^8n^{-3/2})$$

and also,

$$dz_k = dy/(1 + b_nq_5(\tilde{z}, y) + b_n^2\tilde{q}_5(\tilde{z}, y) + 0(n^{-3/2}(\log n)^8)).$$

Hence on C_n,

$$(1 + n^{-1/2}Q_1(z) + n^{-1}Q_2(z))\phi_V(z)$$
$$= (1 + n^{-1/2}Q_1(\tilde{z}, y) + n^{-1}Q_2(\tilde{z}, y))\phi_V(\tilde{z}, y)$$
$$+ (b_nq_6(\tilde{z}, y) + b_n^2\tilde{q}_6(\tilde{z}, y))\phi_V(\tilde{z}, y) + 0(n^{-3/2}(\log n)^\beta)$$

4

where β depends on the degrees of Q_1 and Q_2. So

$$I = \int_{\{g(z)<v\}\bigcap C_n} (1 + n^{-1/2}Q_1(z) + n^{-1}Q_2(z))\phi_V(z)dz$$

$$= \int_{\{x:\ h(x)<v\}\bigcap r(C_n)} (1 + b_n q_6(x) + b_n^2 \tilde{q}_6(x))\phi_V(x)dx + 0(n^{-3/2}(\log n)^\beta).$$

Now, there exists $n_0 = n_0(M, l_0)$ such that for $n \geq n_0$,

$$\{x: \|x\| \leq C \log n\} \subset r(C_n)$$

and

$$\int_{\{x:\ \|x\|>C \log n\}} (1 + b_n|q_6(x)| + b_n^2|\tilde{q}_6(x)|)\phi_V(x)dx = 0(n^{-3/2}).$$

Further

$$\int_{z \notin C_n} (1 + n^{-1/2}Q_1(z) + n^{-1}Q_2(z))\phi_v(z)dz = 0(n^{-3/2}).$$

Thus, using the properties of Fourier-Stieljes transform,

$$\int_{\{z:\ g(z)<v\}} (1 + n^{-1/2}Q_1(z) + n^{-1}Q_2(z))\phi_V(z)dz$$

$$= \int_{\{x:h(x)<v\}} (1 + b_n q_6(x) + b_n^2 \tilde{q}_6(x))\phi_{V'Vl}(y)dy + 0(n^{-3/2}(\log n)^\beta)$$

$$= \int_{-\infty}^{u} (1 + b_n p_1(y) + b_n^2 p_2(y))\phi(y)dy + 0(n^{-3/2}(\log n)^\beta)$$

where p_i's have the properties mentioned in the lemma.

We first state the simplest version of our main theorem. This will illustrate the ideas involved in the general case.

<u>Result 2.3:</u> Let $F, F_n, \underset{\sim}{X}, \underset{\sim}{X}^*$ be as in the introduction.

Let

$$T(\underset{\sim}{X}, F) = n^{1/2}(K(\overline{X}_n) - K(\mu))/\sigma$$

$$T^*(\underset{\sim}{X}^*, F_n) = n^{1/2}(K(\overline{X}_n^*) - K(\overline{X}_n))/\sigma_n$$

where

$$K(x) = l(\mu)'x, \quad (l(\mu) \neq 0)$$

$$\sigma^2 = l(\mu)'\Sigma l(\mu)$$

$$\sigma_n^2 = l(\overline{X}_n)'\Sigma_n l(\overline{X}_n)$$

and Σ and Σ_n are respectively the population and sample dispersion matrices. Further assume that

$$\int \|x\|^{12} dF(x) < \infty \quad \dots \quad \dots \quad \dots \quad \dots \tag{2.1}$$

5

$$\lim_{\|t\|\to\infty} \left| \int e^{it'x} dF(x) \right| < 1 \quad \ldots \quad \ldots \quad \ldots \tag{2.2}$$

Then $\sup_{0\le\alpha\le 1} |P(T(\underset{\sim}{X},F) < H_n^{*^{-1}}(\alpha)) - \alpha| = 0(n^{-1}(\log n)^{1/2})$ where H_n^* is the distribution function of $T^*(\underset{\sim}{X}^*, F_n)$.

Remark 2.4: The condition (2.2) (the so called Cramer's condition) is needed to ensure the existence of Edgeworth expansion for the distribution of T. Without this, a sharp comparison of H_n and H_n^* seems impossible. Finiteness of the fourth moment is needed to get $0(n^{-1})$ expansion for H_n. The extra eight moments are needed to tackle the sample moment appearing in the expansion of H_n^*. This will be clear from the proof. This moment condition will be strengthened when Σ is replaced by its estimate. Note that the conclusion of Result 2.3 is an unconditional probability statement about T. This is an improvement over bootstrap results which hold "almost surely". Further, the result holds uniformly over α, $0 \le \alpha \le 1$.

Proof of Result 2.3: By the Cramer's condition and finiteness of fourth moment, the following expansion is valid for H_n.

$$H_n(x) = \Phi(x) + n^{-1/2}\int_{-\infty}^{x} p(y,F)\phi(y)dy + n^{-1}\int_{-\infty}^{x} p_1(y,F)\phi(y)dy + 0(n^{-1})$$

where $p(\cdot,F)$ and $p_1(\cdot,F)$ are polynomials whose coefficients are polynomials in the moments of F of order ≤ 3 and ≤ 4 respectively. A proof of this can be found in Bhattacharya and Ghosh (1978). The $0(n^{-1})$ term depends on the fourth moment in such a manner, that we can write

$$H_n(x) = \Phi(x) + n^{-1/2}\int_{-\infty}^{x} p(y,F)\phi(y)dy + n^{-1}\epsilon_n(x)$$

where ϵ_n is uniformly bounded by a constant dependent on $M = \int \|x\|^4 dF(x)$. The distribution F_n does not satisfy the Cramer's condition so the above result cannot be applied directly to get an expansion for H_n^*. However, the convergence of the characteristic function of F_n to that of F is "nice enough". This is exploited by Babu and Singh (1984a) who obtain an Edgeworth expansion for H_n^*. Their results contain the proof of the validity of the following expansion for H_n^*.

$$H^*(x) = \Phi(x) + n^{-1/2}\int_{-\infty}^{x} p(y,F_n)\phi(y)dy + n^{-1}\epsilon_n^*(x)$$

where $\sup_x |\epsilon_n^*(x)| \le g(n^{-1}\sum_{i=1}^{n} \|X_i\|^4)$ and g is a smooth continuous function. Thus

$$\sup_x |H_n(x) - H_n^*(x)| \le n^{-1/2}\sup_x |\int_{-\infty}^{x}[p(y,F) - p(y,F_n)]\phi(y)dy|$$
$$+ n^{-1} + n^{-1}\sup_x |\epsilon_n^*(x)|.$$

Note that

$$\sup_x \int_{-\infty}^x |p(y,F) - p(y,F_n)| \phi(y) dy \le C \left| \int \|z\|^3 d(F_n - F) dz \right|$$

$$= C \left| n^{-1} \sum_{i=1}^n (\|X_i\|^3 - E_F(\|X_i\|^3)) \right|.$$

Note that $Y_i = \|X_i\|^3$ are i.i.d. random variables with $E|Y_i|^{2+\delta} < \infty (\delta > 0)$. Thus by moderate deviation results of Michel (1976),

$$P\{ |n^{-1/2} \sum_{i=1}^n (Y_i - E(Y_i))| \ge \delta (\log n)^{1/2} \} \le C n^{-\delta/2} (\log n)^{-1/2} \quad \ldots \quad \ldots \quad \ldots \quad (2.3)$$

(This is where we need $\delta = 2$ and hence 12th moment).

On the other hand,

$$P\{ n^{-1} \sum_{i=1}^n \|X_i\|^4 \ge 2E_F(\|X_i\|^4) \} \le n^{-1} \frac{[E(\|X_1\|^4)]^2}{V(\|X_1\|^4)}$$

$$\le C n^{-1} \quad \ldots \quad \ldots \quad \ldots \quad (2.4)$$

Combining the estimates (2.3) and (2.4), we get, $P\{ n\|H_n - H_n^*\| \ge C(\log n)^{1/2} \} \le C n^{-1}$. Note that $j(H_n) = 0(n^{-1})$. Hence application of Lemma 2.1 proves the result.

Remark 2.5: A completely different approach will perhaps be needed to get rid of the $(\log n)^{1/2}$ factor. Also note that (2.3) is equivalent to $E|Y_i|^{2+\delta'} < \infty \quad \forall \ 0 < \delta' < \delta$. In view of this the moment condition cannot be improved.

The next result deals with non linear statistics. However we still keep the "known variance" assumption.

Result 2.6: Let F, F_n, X, X^* be as in Result 2.3 except that K is not necessarily linear. Suppose K is differentiable with the vector of first partials at μ as $l(\mu) \ne 0$. Assume all other conditions of Result 2.3. We have the following.

(i) If K is thrice continuously differentiable then

$$\sup_{0 \le \alpha \le 1} |P(T(X,F) < H_n^{*^{-1}}(\alpha)) - \alpha| = 0(n^{-1}(\log n)^{3/2})$$

(ii) If K is four times continuously differentiable then the above bound is $0(n^{-1}(\log n)^{1/2})$.

Proof: We first assume K is thrice continuously differentiable.

7

Let

$$Z_n^* = n^{1/2}(\overline{X}_n^* - \overline{X}_n)$$
$$Z_n = n^{1/2}(\overline{X}_n - \mu).$$

On $\|Z_n^*\| \leq C(\log n)^{1/2}$, $\|Z_n\| \leq C(\log n)^{1/2}$ and $|\sigma_n - \sigma| \leq Cn^{-1/2}(\log n)^{1/2}$,

$$\sigma_n T^*(\underset{\sim}{X}^*, F_n) = l(\overline{X}_n)'Z_n^* + n^{-1/2}Z_n^{*'}L(\overline{X}_n)Z_n^* + 0(n^{-1/2}(\log n)^{3/2}) \quad \ldots \quad (2.5)$$

$$\sigma T(\underset{\sim}{X}, F) = l(\mu)'Z_n + n^{-1/2}Z_n'L(\mu)Z_n + 0(n^{-1/2}(\log n)^{3/2}) \quad \ldots \quad (2.6)$$

By moderate deviation result,

$$P(\|Z_n\| \geq C(\log n)^{1/2}) = 0(n^{-1}) \text{ and}$$

$$P(|\sigma_n - \sigma| \geq Cn^{-1/2}(\log n)^{1/2}) = 0(n^{-1}).$$

We claim that

$$\Delta_n = P^*(\|Z_n^*\| \geq C(\log n)^{1/2}) = 0(n^{-1}) \text{ w.p. } 1 - 0(n^{-1}).$$

In what follows, β_n will denote any quantity which is bounded on the set where the third moment is bounded (which happens with probability $1 - 0(n^{-1})$). Note that from the results of Babu and Singh (1984a) (see Theorem 2)

$$\left|\Delta_n - \int_{\{\|x\|\geq C(\log n)^{1/2}\}} (1 + n^{-1/2}p(x, F_n))d\Phi_{\Sigma_n}(x)dx\right| \leq n^{-1}\beta_n.$$

By moderate deviation results,

$$P\{\|\Sigma - \Sigma_n\| \geq Cn^{-1/2}(\log n)^{1/2}\} = 0(n^{-1}).$$

The moments appearing in $p(y, F_n)$ are bounded w.p. $1 - 0(n^{-1})$. Thus it suffices to show that

$$\left|\int_{\{\|x\|\geq C(\log n)^{1/2}\}} (1 + n^{-1/2}p(x))\phi_\Sigma(x)dx\right| \leq Cn^{-1}$$

for any polynomial p. This can be shown very easily. So the claim is proved.

Combining all these estimates we can say that

$$T(\underset{\sim}{X}^*, F_n) = Y_n(Z_n^*)\sigma_n^{-1} + R_n \text{ w.p. } 1 - 0(n^{-1})$$

where

$$Y_n(Z_n^*) = l'(\overline{X}_n)Z_n^* + n^{-1/2}Z_n^{*'}L(\overline{X}_n)Z_n^*$$

and

$$R_n = 0(n^{-1}(\log n)^{3/2}).$$

Denote the distribution function of Z_n^* by H_n^*. With probability $1 - 0(n^{-1})$,

$$P^*(T(\underset{\sim}{X}^*, F_n) \leq x) = \int_{\{z: \ Y_n(z)+R_n \leq \sigma_n x\}} dH_n^*(z)$$

$$= \int_{\{z: \ Y_n(z)+R_n \leq \sigma_n x\}} (1 + n^{-1/2} p(z, F_n)) \phi_{\Sigma_n}(z) dz + n^{-1} \beta_n$$

where the last equality follows from the expansion of the distribution function of Z_n^* given in Babu and Singh (1984a).

Note that w.p. $1 - 0(n^{-1})$,

$$\int_{\{z: \ Y_n(z)+R_n \leq \sigma_n x\}} \phi_{\Sigma_n}(z) dz = \int_{\{z: \ Y_n(z) \leq \sigma_n x\}} \phi_{\Sigma_n}(z) dz + n^{-1} (\log n)^{3/2} \beta_n$$

and

$$\int_{\{z: \ Y_n(z)+R_n \leq \sigma_n x\}} n^{-1/2} p(z, F_n) \phi_{\Sigma_n}(z) dz$$

$$= \int_{\{z: \ Y_n(z) \leq \sigma_n x\}} n^{-1/2} p(z, F_n) \phi_{\Sigma_n}(z) dz + n^{-3/2} (\log n)^{3/2} \beta_n$$

Thus with probability $1 - 0(n^{-1})$,

$$P^*(T(\underset{\sim}{X}^*, F_n) \leq x) = \int_{\{z: \ Y_n(z) \leq \sigma_n x\}} (1 + n^{-1/2} p(z, F_n)) \phi_{\Sigma_n}(z) dz + n^{-1} (\log n)^{3/2} \beta_n$$

which by Lemma 2.2 (i)

$$= \int_\infty^x (1 + b_n^* p_1(y, F_n)) \phi(y) dy + n^{-1} (\log n)^{3/2} \beta_n.$$

A similar expansion holds for $P(T(\underset{\sim}{X}, F) \leq x)$:

$$P(T(\underset{\sim}{X}, F) \leq x) = \int_{-\infty}^x (1 + b_n p_1(y, F)) \phi(y) dy + 0(n^{-1} (\log n)^{3/2}).$$

Note that b_n, b_n^* are smooth functions of the moments. The coefficients in the polynomial p_1 are also so. Hence proceeding as in the last part of the proof of Result 2.3,

$$P\{n||H_n - H_n^*|| \geq C(\log n)^{3/2}\} \leq Cn^{-1}.$$

The proof of (i) now follows from Lemma 2.1.

Now assume that K is four times continuously differentiable. T and T^* are now expanded upto three terms and the remainders are $0(n^{-3/2} (\log n)^\alpha)$. The proof now is exactly as before except that we now have to use Lemma 2.2 (ii). This finishes the proof of Result 2.6.

9

We now discuss Studentized versions of the statistics of Result 2.3 and 2.6.

Let

$$t(\underset{\sim}{X}, F) = n^{1/2}(K(\overline{X}_n) - K(\mu))/\nu(n^{-1}\sum_{i=1}^{n}\lambda(X_i))$$

$$t^*(\underset{\sim}{X}^*, F) = n^{1/2}(K(\overline{X}_n^*) - K(\overline{X}_n))/\nu(n^{-1}\sum_{i=1}^{n}\lambda(X_i^*))$$

where $\lambda(\cdot) = (\lambda_1(\cdot), \ldots, \lambda_g(\cdot))$ is a continuous function from $\mathbb{R}^k$ to $\mathbb{R}^g$. $\nu(\cdot)$ is a real valued function on $\mathbb{R}^g$ such that

$$\nu(E_F\lambda(X_1)) = (l'(\mu)\Sigma l(\mu))^{1/2} = \sigma$$
$$\nu(E_{F_n}\lambda(X_1^*)) = (l'(\overline{X}_n)\Sigma_n l(\mu_n))^{1/2} = \sigma_n.$$

As before l' is the vector of first derivatives of K, Σ and Σ_n are the population and sample covariance matrices of X_i.

Let $L(X_i)$ be a linearly independent subcollection of $(X_i, \lambda(X_i))$, such that all the elements of $(X_i, \lambda(X_i))$ can be expressed as linear combinations of those of $L(X_i)$. The following theorem gives bootstrap accuracy of t. As before the accuracy depends on smoothness of K (and ν).

<u>Theorem 2.7</u>: Suppose that the distribution $\tilde{F}$ of $L(X_1)$ satisfies (2.1) and (2.2).

(i) If K is thrice continuously differentiable and ν is twice continuously differentiable then

$$\sup_{0\leq\alpha\leq 1} |P(t(\underset{\sim}{X}, F) < H_n^{*^{-1}}(\alpha)) - \alpha| = 0(n^{-1}(\log n)^{3/2}) \quad \ldots \qquad (2.7)$$

(ii) If K is four times continuously differentiable and ν is thrice continuously differentiable then the rate in (2.7) is $0(n^{-1}(\log n)^{1/2})$.

<u>Proof</u>: Let

$$v_n^2 = \nu(n^{-1}\sum_{i=1}^{n}\lambda(X_i^*))$$

$$\eta_n = n^{-1/2}\sum_{i=1}^{n}(L(X_i) - E_F(L(X_i)))$$

$$Z_n = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu)$$

$$Y_n = n^{-1/2}\sum_{i=1}^{n}(\lambda(X_i) - E_F(\lambda(X_1)))$$

10

and η_n^*, Z_n^* and Y_n^* are their corresponding bootstrap versions. e.g.,

$$\eta_n^* = n^{-1/2} \sum_{i=1}^n (L(X_i^*) - E_{F_n}(L(X_1^*))).$$

Let

$$C_n = \{\|\eta_n\| + \|\eta_n^*\| \le C(\log n)^{1/2}\}.$$

We will prove (i) first.

On C_n, T and T^* have the expansions (2.5) and (2.6) and

$$\sigma_n - \sigma = n^{-1/2} Y_n' \nu^{(1)}(E_F(\lambda(X_1))) + 0(n^{-1}\log n)$$
$$v_n - \sigma_n = n^{-1/2} Y_n^{*'} \nu^{(1)}(E_{F_n}\lambda(X_1^*))) + 0(n^{-1}\log n)$$

where $\nu^{(1)}$ is the vector of first partials of ν. Combining these expansions we can write,

$$t(\underset{\sim}{X}, F) = U'\eta_n + n^{-1/2}\eta_n' B\eta_n + 0(n^{-1}(\log n)^{3/2})$$
$$t^*(\underset{\sim}{X}^*, F) = U_n'\eta_n^* + n^{-1/2}\eta_n^{*'} B_n \eta_n^* + 0(n^{-1}(\log n)^{3/2})$$

where U, B are functions of moments of $L(X_1)$ and l and $\nu^{(1)}$. U_n and B_n are same as U and B, with the population moments replaced by the sample moments. Now the proof is as in Result 2.6.

(ii) In this case, T, T^*, $\sigma_n - \sigma$ and $v_n - \sigma_n$ are expanded upto one more term. Thus t and t^* are expanded to three terms with error $0(n^{-1})$. Now the proof is same as in the last part of Result 2.6.

Remark 2.8: (i) The usual choice of $\nu(\cdot)$ is such that

$$\nu(n^{-1}\sum_{i=1}^n \lambda(X_i)) = l'(\overline{X}_n)\Sigma_n l(X_n) \text{ and}$$
$$\nu(n^{-1}\sum_{i=1}^n \lambda(X_i^*)) = l'(\overline{X}_n^*)D_n l(\overline{X}_n^*)$$

where D_n is the dispersion matrix of X_i^*. But in some parametric situations, simpler options are available.

(ii) An inspection of the proof of Theorem 2.7 shows that Edgeworth expansion for the distribution of Z_n is not enough because of the product factors coming from T and σ_n. These involve linear and quadratic terms of $L(X_1)$ and as we have seen in Result 2.3, we need the 12th moment of $L(X_1)$. For the Student's t statistic, $L(X_1) = (X_1, X_1^2)$ and hence we require $E|X_1|^{24} < \infty$. We believe that some alternate approach will reduce this moment condition.

11

References

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Statistics*, **13**, 116–132.

Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit. *Sankhyā*, **46**, 85–93.

Babu, G. J. and Singh, K. (1983). Nonparametric inference on means using bootstrap. *Ann. Statistics*, **11**, 999–1003.

Babu, G. J. and Singh, K. (1984a). On one term Edgeworth correction by Efron's bootstrap. *Sankhyā* **A 46**, 219–232.

Babu, G. J. and Singh, K. (1984b). Asymptotic representations related to jackknifing and bootstrapping *L*-statistics. *Sankhyā* **A 46**, 195–206.

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

Bickel, P. J. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statistics* **9**, 1196–1217.

Efron, B. (1979). Bootstrap methods: another look at Jackknife. *Ann. Statistics*, **7**, 1–26.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, **72**, 45–58.

Efron, B. (1987). Better bootstrap confidence intervals. *Jour. Amer. Stat. Assoc.* **82**, No. 397, 171–185.

Michel, R. (1976). Non-uniform central limit bounds with applications to probabilities of deviations. *Ann. Prob.*, **4**, 102–106.

Singh, K. (1981). On asymptotic accuracy of Efron's bootstrap. *Ann. Statistics*, **9**, 1187–1195.