ON CONVERGENCE OF DENSITIES OF TRANSLATION AND SCALE STATISTICS

Dennis D. Boos

Inst. of Statistics Mimeo Series #1625

MIMEOGRAPH SERIES #1625

ON CONVERGENCE OF DENSITIES OF TRANSLATION AND SCALE STATISTICS

ABSTRACT

Convergence of densities leads directly to convergence of their distribution functions via Scheffé's theorem. This paper is concerned with the converse: what are sufficient conditions to obtain convergence of densities from convergence of distribution functions? A general lemma is given and results are obtained for translation and scale statistics with applications to bootstrap estimates of densities and convergence of Bayes posteriors.

Key words and phrases. Convergence of densities, translation statistics, scale statistics, bootstrap, Bayes posteriors.

1. Introduction

When does convergence in distribution imply convergence of the associated densities? Specifically, let $T_n = (a_{1n}(T_{1n} - b_{1n}), \cdots, a_{kn}(T_{kn} - b_{kn}))$ be a standardized random vector in R^k which converges weakly to a distribution having density (Radon - Nikodym derivative) g with respect to (wrt) Lebesgue measure μ on R^k. If T_n has density g_n wrt μ, what are sufficient conditions for g_n to converge to g, say pointwise a.e. μ? Lemma 1 below gives one set of such conditions and Section 2 verifies these conditions for certain translation and scale statistics. However, the results are very modest and it is hoped that this paper will stimulate more general methods. Apparently the present literature does not address this problem except for a few fragmented local limit theorems. Section 3 uses Lemma 1 and the ideas of Section 2 to show that bootstrap *estimates* of densities of statistics converge as expected. Section 4 discusses convergence of some Bayes posteriors which originally motivated the problem.

Let G_n and G denote the distribution functions (dfs) of g_n and g and let "$\Rightarrow$" stand for weak convergence. If $G_n \Rightarrow G$, a sufficient condition for $g_n(x) \to g(x)$ a.e. μ is that each subsequence $g_{n'}$ have a further subsequence $g_{n''}$ converging to some density g^*. In that case, Scheffé's theorem (Serfling, 1980, p. 17) yields $G_{n''} \Rightarrow G^*$ which would contradict $G_n \Rightarrow G$ if $g^* \neq g$ on a set of positive μ measure. However, the existence of convergent subsequences is not easy to verify in the absence of a metric space. (In the topology of pointwise convergence g_n is compact if $\{g_n\}$ is is closed and $\{g_n(x)\}$ is bounded for each x. Unfortunately, this topology is not metrizable here.) Thus, it is convenient to use uniform convergence on compacts for which the Ascoli theorem is available (c.f., Royden, 1968, p. 179). This leads to Lemma 1 and (1.3). If the equicontinuity is uniform, then we can extend the result to uniform convergence on R^k, i.e., wrt the norm $\sup_x |g_n(x) - g(x)| \equiv \|g_n - g\|_\infty$. The proof of this extension was suggested by Chris Klaassen.

AMS 1980 *subject classifications*, Primary 62E20; secondary 62G05.

LEMMA 1. Suppose that G_n and G have continuous densities g_n and g wrt μ on R^k. If $G_n \Rightarrow G$ and

(1.1) $\sup_n |g_n(x)| \le M(x) < \infty$, each $x \in R^k$,

and

(1.2) $\{g_n\}$ is equicontinuous, i.e., for each x and $\varepsilon > 0$ there exists $\delta(x,\varepsilon)$ and $n(x,\varepsilon)$ such that $|x-y| < \delta(x,\varepsilon)$ implies that $|g_n(x) - g_n(y)| < \varepsilon$ for all $n \ge n(x,\varepsilon)$,

then for any compact subset C of R^k

(1.3) $$\sup_{x \in C} |g_n(x) - g(x)| \to 0 \quad \text{as} \quad n \to 0 .$$

If $\{g_n\}$ is uniformly equicontinuous, i.e., $\delta(x,\varepsilon) = \delta(\varepsilon)$ and $n(x,\varepsilon) = n(\varepsilon)$ in (1.2) do not depend on x, then

(1.4) $$\sup_{x \in R^k} | g_n(x) - g(x)| = \|g_n - g\|_\infty \to 0 \quad \text{as} \quad n \to \infty.$$

PROOF. Conditions (1.1) and (1.2) are exactly what the classical Ascoli theorem requires for $\{g_n\}$ to be compact wrt the topology of uniform convergence on compacts. Thus, if $g_{n'}$ is a subsequence of g_n, then there is a further subsequence $g_{n''}$ which converges uniformly on each compact subset of R^k to some g^*. Scheffé's theorem then shows that $g^* = g$ a.e. μ, and since they are both continuous $g^*(x) = g(x)$, each $x \in R^k$. So, for each compact set C

(1.5) $$\sup_{x \in C} |g_{n''}(x) - g(x)| \to 0 .$$

It should be clear from the following argument how (1.5) yields (1.3). Now, suppose that (1.4) is false. Then, there must exist an $\varepsilon > 0$ and a subsequence n' of n and a sequence $x_{n'}$ such that

(1.6) $$|g_{n'}(x_{n'}) - g(x_{n'})| > \varepsilon \qquad \text{for all } n' .$$

If the $x_{n'}$ are bounded, then there exists some C which contain all the $x_{n'}$ and (1.6) contradicts (1.5). If the $x_{n'}$ are not bounded, then there is a subsequence $x_{n''}$ such that at least one coordinate of $x_{n''}$ is tending to $\pm\infty$. Since g is a density we have $g(x_{n''}) \to 0$ as $n'' \to \infty$ and thus by (1.6) $g_{n''}(x_{n''}) \geq \varepsilon/2$ for all n'' sufficiently large, say all $n'' \geq n_0$. Let $\delta = \delta(\varepsilon/4)$ be such that $|x_{n''} - y| < \delta$ implies $|g_{n''}(x_{n''}) - g_{n''}(y)| < \varepsilon/4$ for all $n \geq n(\varepsilon/4)$. Then

$$(1.7) \qquad G_{n''}(x_{n''} + \delta) - G_{n''}(x_{n''} - \delta) = \int_{x_{n''} - \delta}^{x_{n''} + \delta} g_{n''}(y)dy \geq (2\delta)^k \varepsilon/4$$

for all $n'' \geq \max(n_0, n(\varepsilon/4))$ since $g_{n''}(y) \geq \varepsilon/4$ for all y in the δ neighborhood of $x_{n''}$ (let $|\cdot|$ on R^k be the maximum of the coordinates). But $G_{n''}(x_n + \delta) - G_{n''}(x_{n''} - \delta) \leq 2\|G_n - G\|_\infty + G(x_{n''} + \delta) - G(x_{n''} - \delta)$, and $\|G_n - G\|_\infty \to 0$ by Polya's theorem in R^k and $G(x_{n''} + \delta) - G(x_{n''} - \delta) = g(x^*_{n''})(2\delta)^k$, where $x^*_{n''} \in (x_{n''} - \delta, x_{n''} + \delta)$. Since at least one coordinate of $x_{n''} \to \pm\infty$, the same is true for $x^*_{n''}$. Thus, $g(x^*_{n''}) \to 0$ as $n'' \to \infty$ and (1.7) is contradicted. □

2. Translation and scale statistics.

a. Univariate statistics.

A statistic $T(X) = T(X_1,\cdots,X_n)$ is a *translation* statistic t if $T(x+a) = T(x)+a$ for all real a and vectors $x = (x_1,\cdots, x_n)$. Let $X_1,\cdots,X_n$ be independent and identically distributed (iid) observations with common density $f(x)$. Then, using the translation property we have

$$G_n(y+t) = P(T(X) \le y+t) = \int I(T(x) \le y+t) \prod_{i=1}^{n} f(x_i)dx_1\cdots dx_n$$
$$= 1 - \int I(T(x) > y) \prod_{i=1}^{n} f(x_i+t)dx_1\cdots dx_n$$
$$= 1 - \int I(T(x) > y)\exp\{\sum_{i=1}^{n} \log f(x_i+t)\}dx_1\cdots dx_n .$$

Now, to get derivatives of $G_n(y)$ we need only justify the interchange of operations

(2.1) $$\frac{d^k}{dy^k} G_n(y) = \frac{d^k}{dt^k} G_n(y+t)\Big|_{t=0} = -\int I(T(x) > y)\frac{d^k}{dt^k}\exp\{\sum_{i=1}^{n} \log f(x_i+t)\}\Big|_{t=0} dx_1\cdots dx_n .$$

Klaassen (1982) introduced this approach and showed that if f is absolutely continuous with $\int|f'(x)|dx < \infty$, then

(2.2) $$g_n(y) = \int I(T(x) > y)[-\sum_{i=1}^{n} \frac{f'}{f}(x_i)] \prod_{i=1}^{n} f(x_i)dx_1\cdots dx_n .$$

(For notational convenience, first and second derivatives of f are often written as f' and f'' and $d^k f(x)/dx^k$ is sometimes denoted $f^{(k)}(x)$. If G_n is a df then g_n is its first derivative.) In a similar fashion one can show that if in addition f' is absolutely continuous with $\int|f''(x)|dx < \infty$, then

(2.3) $$g_n'(y) = \int I(T(x) > y)[-\sum_{i=1}^{n} \frac{f''}{f}(x_i) - \sum_{i\ne j} \frac{f'}{f}(x_i)\frac{f'}{f}(x_j)] \prod_{i=1}^{n} f(x_i)dx_1\cdots dx_n.$$

It appears that generally one needs $f^{(k-1)}$ to be absolutely continuous with $\int|f^{(k)}(x)|dx < \infty$ in order to justify (2.1). Unfortunately, the expressions tend to get messy for $k > 2$.

We can get similar representations for the logarithm of positive *scale* statistics $S(X)$, i.e., those satisfying $S(ax) = aS(x) > 0$ for all $a > 0$ and $x = (x_1, \cdots x_n)$. Then,

$$(2.4) \qquad H_n(y+t) = P(\log S(x) \le y+t) = 1 - \int I(\log S(x) > y) \exp\{\sum_{i=1}^{n} \log(e^t f(e^t x_i))\} dx_1 \cdots dx_n .$$

If f is absolutely continuous with $\int |xf'(x)| dx < \infty$, then

$$(2.5) \qquad h_n(y) = \int I(\log S(x) > y) \sum_{i=1}^{n} (-1 - x_i \frac{f'}{f}(x_i)) \prod_{i=1}^{n} f(x_i) dx_1 \cdots dx_n .$$

Obviously, higher order derivatives may be taken under sufficient regularity conditions. A similar approach is also possible for two-sample shift statistics and ratios of positive scale statistics.

The first theorems can now be given. Let $I(f) = \int_{-\infty}^{\infty} [f'(x)/f(x)]^2 f(x) dx$, $I_1(f) = \int_{-\infty}^{\infty} [-1 - xf'(x)/f(x)]^2 f(x) dx$, and let $\Phi(x)$ and $\phi(x)$ be the standard normal df and density. F is the df associated with f and $T(F)$ is the target "parameter" of T_n.

THEOREM 1. Let $X_1, \cdots, X_n$ be independent with common density $f(x)$ on R^1. Let T_n be a translation statistic such that $G_n(y) = P(n^{\frac{1}{2}}[T_n - T(F)] \le y) \to \Phi(y/\sigma)$, each y. If f is absolutely continuous and $I(f) < \infty$, then with $g(x) = \sigma^{-1}\phi(x/\sigma)$ we have

$$\| g_n - g \|_\infty \to 0 \qquad \text{as} \qquad n \to \infty .$$

THEOREM 2. Let $X_1, \cdots, X_n$ be independent with common density $f(x)$ on R^1. Let S_n be a positive scale statistic such that $H_n(y) = P(n^{\frac{1}{2}}[\log S - \log S(F)] \le y) \to \Phi(y/\sigma)$, each y. If f is absolutely continuous and $I_1(f) < \infty$, then with $h(x) = \sigma^{-1}\phi(x/\sigma)$ we have

$$\| h_n - h \|_\infty \to 0 \qquad \text{as} \qquad n \to \infty .$$

PROOFS. Since both proofs are virtually the same, only the first will be given. From $I(f) < \infty$ via Cauchy-Schwarz we get $\int |f'(x)| dx < \infty$ and thus (2.2) with $g_n(y)$ replaced by $n^{-\frac{1}{2}} g_n(n^{-\frac{1}{2}} y + T(F))$ can be written as

$$g_n(y) = E[n^{-\frac{1}{2}} \sum_{i=1}^{n} - \frac{f'}{f}(x_i)] I(n^{\frac{1}{2}}(T_n - T(F)) > y) .$$

One application of the Cauchy-Schwarz inequality yields $g_n(y) \leq [I(f)]^{\frac{1}{2}}$ and another gives

$$|g_n(x) - g_n(y)| \leq [I(f)]^{\frac{1}{2}} |G_n(x) - G_n(y)|^{\frac{1}{2}} \leq I(f)|x - y|^{\frac{1}{2}} .$$

Hence, $\{g_n\}$ is uniformly bounded and uniformly equicontinuous so that Lemma 1 applies. □

The analogous results for two-sample shift statistics and ratios of positive scale statistics should be clear. The next result relates to the derivative of the density in Theorem 1. Here the proof is different since Scheffé's theorem is not available to justify "$g_n' \to g'$ implies $g_n \to g$."

THEOREM 3. Assume the conditions of Theorem 1. In addition, suppose that f' is absolutely continuous and $\int |f''(x)| dx < \infty$. If $(n^{\frac{1}{2}}[T_n - T(F)], n^{-\frac{1}{2}} \sum_1^n f'(X_i)/f(X_i))$ converges in distribution to a bivariate normal $(Z_1, Z_2)^T$ with means 0 and variances σ^2 and $I(f)$, then

$$g_n'(y) \to \frac{1}{\sigma^2} \phi'(\frac{y}{\sigma}), \text{ each } y, \quad \text{as } n \to \infty.$$

PROOF. (2.3) with $g_n'(y)$ replaced by $n^{-1} g_n'(n^{-1}y + T(F))$ can be written as

$$g_n'(y) = E\{- \frac{1}{n} \sum_{i=1}^{n} \frac{f''}{f}(X_i) + \frac{1}{n} \sum_{i=1}^{n} [\frac{f'}{f}(X_i)]^2 - [n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{f'}{f}(X_i)]^2\} I(n^{\frac{1}{2}} (T_n - T(F)) > y).$$

The quantity inside this expectation is uniformly integrable and converges in distribution to the random variable $I(f) - Z_2^2 I(Z_1 > y)$. Then, Theorem 1.4A, p. 14 of Serfling (1980), yields $g_n'(y) \to E[I(f) - Z_2^2 I(Z_1 > y)]$. A simple calculation shows that this latter expectation has the desired value. □

Remark 2.1. The joint convergence in the hypothesis is usually shown by justifying the expansion $n^{\frac{1}{2}}[T_n - T(F)] = n^{-\frac{1}{2}} \sum_1^n \psi(X_i) + R_n$, where $R_n \xrightarrow{P} 0$ and $E\psi(X_i) = 0$ and $\text{Var}\psi(X_i) = \sigma^2$.

b. Joint densities.

The methods of the last subsection are not ideally suited for joint densities. The next three theorems show the kinds of results one can get. The first is for application to studentized statistics, the second is for regression, and the third is for bivariate location. The related convergence in distribution results are often made much easier by Slutsky's theorem and the Cramér-Wold device. Let $\Phi(y_1,y_2;0,\Sigma)$ and $\phi(y_1,y_2;0,\Sigma)$ be the bivariate normal df and density with mean 0 and nonsingular covariance matrix Σ having diagonal elements σ_1^2 and σ_2^2.

THEOREM 4. Let $X_1,\cdots X_n$ be independent with common density $f(x)$ on R^1. Let T_n be a translation statistic and S_n a scale statistic such that $G_n(y_1,y_2) = P(n^{\frac{1}{2}}[T_n - T(F)] \le y_1,\ n^{\frac{1}{2}}[\log S_n - \log S(F)] \le y_2) \to \Phi(y_1,y_2;0,\Sigma)$. If f and f' are absolutely continuous such that $I(f) < \infty$, $I_1(f) < \infty$, and $\int|f'(x)/f(x)|^{1+\varepsilon}f(x)dx < \infty$ and $\int|xf''(x)/f(x)|^{1+\varepsilon}f(x)dx < \infty$ for some $\varepsilon > 0$, then

$$\begin{aligned} g_n(y_1,y_2) &= \int I(n^{\frac{1}{2}}[T_n - T(F)] \le y_1,\ n^{\frac{1}{2}}[\log S_n - \log S(F)] \le y_2) \\ &\times \{\frac{1}{n}\sum_{i=1}^{n}(2\frac{f'}{f}(x_i) + x_i\frac{f''}{f}(x_i)) + \frac{1}{n}\sum_{i\neq j}(1 + x_i\frac{f'}{f}(x_i))(\frac{f'}{f}(x_j))\} \\ &\times \prod_{i=1}^{n} f(x_i)dx_1\cdots dx_n, \end{aligned}$$

and for $g(y_1,y_2) = \phi(y_1,y_2;0,\Sigma)$ we have

$$\| g_n - g\|_\infty \to 0 \quad \text{as} \quad n \to \infty .$$

COROLLARY. The density of the studentized statistic $V_n = n^{\frac{1}{2}}[T_n - T(F)]/S_n$, say $g_{V_n}(y)$, converges in $L_1(-\infty,\infty)$ to $g_V(y) = S(F)\phi(S(F)y/\sigma_1)$.

PROOF. Similar to (2.3) the representation for g_n follows from the hypotheses of the theorem. Uniform boundedness and equicontinuity follow by moment calculations where the $1+\varepsilon$ enters when using Hölder's inequality in verifying the equicontinuity. For the Corollary, let $W_n = n^{\frac{1}{2}}[\log S_n - \log S(F)]$. Using straightforward transformation, the density of (V_n, W_n) is

$$g_{V_n,W_n}(y_1,y_2) = S(F)\exp\{n^{-\frac{1}{2}}y_2\}g_n(S(F)\exp\{n^{-\frac{1}{2}}y_2\}y_1,y_2) \ .$$

Using the equicontinuity of g_n and the convergence of g_n from Theorem 4, we have that g_{V_n,W_n} converges pointwise to $g_{V,W}(y_1,y_2) = S(F)\phi(S(F)y_1,y_2;0,\Sigma)$. Thus by Scheffé's theorem g_{V_n,W_n} converges to $g_{V,W}$ in $L_1(R^2)$. Then

$$\int_{-\infty}^{\infty}|g_{V_n}(y) - g_V(y)|dy = \int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty}[g_{V_n,W_n}(y,z) - g_{V,W}(y,z)]dz\right|dy$$

$$\le \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g_{V_n,W_n}(y,z) - g_{V,W}(y,z)|dz\,dy \to 0 \ . \ \square$$

Consider the familiar linear model $Y = X\beta + \varepsilon$, where $\beta^T = (\beta_1,\cdots,\beta_k)$ are the parameters and $\varepsilon = (\varepsilon_1,\cdots,\varepsilon_n)^T$ are independent errors with common density $f(x)$ on R^1. Let us say that $\hat\beta$ is a translation estimator if $\hat\beta(Y + Xt, X) = \hat\beta(Y,X) + t$ for all (Y,X) and $t = (t_1,\cdots,t_k)^T$. In that case a joint representation for any two components is fairly simple since $\hat\beta$ may be viewed as a function of the ε_i (given X and β). Let ℓ and m be integers from $\{1,\cdots k\}$. Then

$$G_n(y_\ell + t_\ell, y_m + t_m) = P(n^{\frac{1}{2}}[\hat\beta_\ell - \beta_\ell] \le y_\ell + t_\ell, n^{\frac{1}{2}}[\hat\beta_m - \beta_m] \le y_m + t_m)$$

$$= \int I(n^{\frac{1}{2}}[\hat\beta_\ell(\varepsilon - x_\ell t_\ell n^{-\frac{1}{2}}) - \hat\beta_\ell] \le y_\ell,\ n^{\frac{1}{2}}[\hat\beta_m(\varepsilon - x_m t_m n^{-\frac{1}{2}}) - \beta_m] \le y_m)\prod_{i=1}^{n} f(\varepsilon_i)d\varepsilon_1\cdots d\varepsilon_n$$

$$= \int I(n^{\frac{1}{2}}[\hat\beta_\ell(\varepsilon) - \beta_\ell] \le y_\ell, n^{\frac{1}{2}}[\hat\beta_m(\varepsilon) - \beta_m] \le y_m)\prod_{i=1}^{n} f(\varepsilon_i + x_{i\ell}t_\ell + x_{im}t_m)d\varepsilon_1\cdots d\varepsilon_n \ ,$$

where x_ℓ is the ℓth column of X and the x_{ij} are the entries of X. If f and f' are absolutely continuous with $\int|f'(x)|dx < \infty$ and $\int|f''(x)|dx < \infty$, then we can justify the interchange of derivative and integration to get

$$g_n(y_\ell, y_m) = \int I(n^{\frac{1}{2}}[\hat{\beta}_\ell - \beta_\ell] \le y_\ell, n^{\frac{1}{2}}[\hat{\beta}_m - \beta_m] \le y_m)$$

$$\times \{\frac{1}{n}\sum_{i=1}^{n} x_{i\ell}x_{im}\frac{f''}{f}(\varepsilon_i) + \frac{1}{n}\sum_{i\neq j} x_{i\ell}x_{jm}\frac{f'}{f}(\varepsilon_i)\frac{f'}{f}(\varepsilon_j)\}\prod_{i=1}^{n} f(\varepsilon_i)d\varepsilon_i\cdots d\varepsilon_n .$$

This representation leads to the following theorem.

THEOREM 5. Under the linear model assumptions above, let $\hat{\beta}$ be a translation estimator of β such that $G_n(y_\ell, y_m) \to \Phi(y_\ell, y_m; 0, \Sigma)$. Suppose that $X^T X/n$ stays bounded and that f and f' are absolutely continuous with $I(f) < \infty$ and $\int |f''(x)/f(x)|^{1+\varepsilon} f(x)dx < \infty$ for some $\varepsilon > 0$, then for $g(y_\ell, y_m) = \phi(y_\ell, y_m; 0, \Sigma)$ we have $||g_n - g||_\infty \to 0$ as $n \to \infty$.

PROOF. The uniform boundedness and equicontinuity are again shown by moment calculations. □

The last result in this subsection is for bivariate densities. The proof is similar to the previous ones. Extension to more dimensions then 2 becomes quite messy.

THEOREM 6. Let $X_1, \cdots, X_n$ be independent with common density $f(x) = f(x_1, x_2)$ on R^2. Let (T_{1n}, T_{2n}) be a translation statistic such that
$G_n(y_1, y_2) = P(n^{\frac{1}{2}}[T_{1n} - T_1(F)] \le y_1, n^{\frac{1}{2}}[T_{2n} - T_2(F)] \le y_2) \to \Phi(y_1, y_2; 0, \Sigma)$.
If f and its first partial derivatives are absolutely continuous with $\int |\partial f(x)/\partial x_1| dx_1 dx_2 < \infty$, $\int |\partial f(x)/\partial x_2| dx_1 dx_2 < \infty$, and $\int |\partial^2 f(x)/\partial x_1 \partial x_2| dx_1 dx_2 < \infty$, then

$$g_n(y_1, y_2) = \int I(n^{\frac{1}{2}}[T_{1n} - T_1(F)] \le y_1,\ n^{\frac{1}{2}}[T_{2n} - T_2(F)] \le y_2)$$

$$\times \{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial z_1 \partial z_2}\log f(z)\Big|_{z=x_i} + [n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial}{\partial z_1}\log f(z)\Big|_{z=x_i}][n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial}{\partial z_2}\log f(z)\Big|_{z=x_i}]\}$$

$$\times \sum_{i=1}^{n} f(x_i)dx_1\cdots dx_n .$$

If further, $E(\partial \log f(z)/\partial z_1|_{z=X_1})^2 < \infty$, $E(\partial \log f(z)/\partial z_2|_{z=X_1})^2 < \infty$, and $E|\partial^2 \log f(z)/\partial z_1 \partial z_2|_{z=X_1}|^{1+\varepsilon} < \infty$ for some $\varepsilon < 0$, then $g_n(y_1,y_2)$ is uniformly bounded and uniformly equicontinuous and with $g(y_1,y_2) = \phi(y_1,y_2;0,\Sigma)$ we have $||g_n - g||_\infty \to 0$ as $n \to \infty$.

3. Bootstrap estimation of densities.

Suppose that we are interested in estimating the density of $n^{\frac{1}{2}}[T_n(X) - T(F)]$ using a random sample $X = (X_1,\cdots,X_n)$ from f. The bootstrap method (Efron, 1979) is to estimate f by say $\hat{f}_n(x) = (nh_n)^{-1}\sum_1^n k((x-X_i)/h_n)$, and then compute the density of $n^{\frac{1}{2}}[T_n(X^*) - T_n(X)]$ (for fixed X) as an estimate of the density of $n^{\frac{1}{2}}[T_n(X) - T(F)]$, where $X^* = (X_1^*,\cdots,X_n^*)$ is a random sample fron $\hat{f}_n$. In practice, one typically generates independent X^* on the computer and estimates the density of $n^{\frac{1}{2}}[T_n(X^*) - T_n(X)]$. Bickel and Freedman (1981) have given some basic theory for concluding that the *distribution* of $n^{\frac{1}{2}}[T_n(X^*) - T_n(X)]$ approximates the distribution of $n^{\frac{1}{2}}[T_n(X) - T(F)]$ in large samples. (Some modifications are required for the theory to hold when the X^* are drawn from $\hat{f}_n$ rather than from the empirical distribution function; see the examples below.) In the next theorem I want to show how to apply the ideas of the last section to get the stronger result concerning convergence of the densities. The theorem will be stated for kernel estimators $\hat{f}_n$ and translation statistics $T_n(X)$. However, similar theorems could be stated for other types of $\hat{f}_n$ and $T_n(X)$. As before, let $G_n(x) = P(n^{\frac{1}{2}}[T_n(X) - T(F)] \le x)$ and $g_n(x) = dG_n(x)/dx$. Let $\hat{F}_n(x) = \int_{-}^{x} \hat{f}_n(y)dy$, $\hat{G}_n(x) = P^*(n^{\frac{1}{2}}[T_n(X^*) - T_n(X)] \le x)$ and $\hat{g}_n(x) = d\hat{G}_n(x)/dx$, where "P*" refers to the conditional probability of X^* given X. Statements "with probability 1" (wp1) refer to the unconditional probability generated by X.

THEOREM 7. Let $X^* = (X_1^*,\cdots,X_n^*)$ be a (conditional) random sample from $\hat{f}_n(x) = (nh_n)^{-1}\sum_1^n k((x - X_i)/h_n)$ where $X = (X_1,\cdots,X_n)$ is a random sample from f(x). Let T_n be a translation statistic.

i) If k is absolutely continuous such that $\int|k'(x)|dx < \infty$, then $\hat{f}_n$ is absolutely continuous, $\int|\hat{f}_n'(x)|dx < \infty$, and

$$\hat{g}_n(y)=\int_{R^n} I(n^{\frac{1}{2}}[T_n(z) - T_n(X)] > y)[n^{-\frac{1}{2}} \sum_{i=1}^{n} - \frac{\hat{f}_n'}{\hat{f}_n}(z_i)] \prod_{i=1}^{n} \hat{f}_n(z_i)dz_1\cdots dz_n .$$

ii) If $I(\hat{f}_n) \le M < \infty$ wpl, all n, and $\hat{G}_n \overset{wpl}{\to} \Phi(y/\sigma)$, then for $g(y) = \sigma^{-1}\phi(y/\sigma)$

$$\|\hat{g}_n - g\|_\infty \to 0 \text{ as } n \to \infty, \text{ wpl} .$$

REMARKS 3.1. The proof of the theorem is trivially like Theorem 1 using (2.2). Virtually all kernels k in use except the uniform satisfy i). The bound on $I(\hat{f}_n)$ is not trivial and will be dealt with elsewhere (see Bhattacharya (1967)). The centering constant $T_n(X)$ in the definition of $\hat{G}_n$ can be replaced by anything which still allows $\hat{G}_n$ to converge. Now we shall examine some typical statistics to which Theorem 7 applies. Each statistic has the functional form $T_n(X) = T(F_n)$, where F_n is the empirical df of X. Let F_n^* be the empirical of df of X^*. For showing $\hat{G}_n(x) \overset{wpl}{\to} \Phi(x/\sigma)$, it is easiest to consider $T_n(X)$ replaced by $T(\hat{F}_n)$, i.e., let $\hat{G}_n(x) = P^*(n^{\frac{1}{2}}[T(F_n^*) - T(\hat{F}_n^*)] \le x)$. The basic approach in Bickel and Freedman (1981) (hereafter called B & F) is to show for almost all sequences $X_1, X_2 \cdots$

(3.1a) $$n^{\frac{1}{2}}[T(F_n^*) - T(\hat{F}_n) - T(\hat{F}_n; F_n^* - \hat{F}_n)] \overset{p^*}{\to} 0$$

(3.1b) $$n^{\frac{1}{2}}[T(\hat{F}_n; F_n^* - \hat{F}_n) - T(F; F_n^* - \hat{F}_n)] \overset{p^*}{\to} 0$$

(3.1c) $$n^{\frac{1}{2}}T(F; F_n^* - \hat{F}_n) \overset{d^*}{\to} N(0, \sigma^2) .$$

Here $T(F;\Delta)$ is the usual linear approximation which reduces to a sum of (conditional) iid random variables when $\Delta = F_n^* - \hat{F}_n$ (see Serfling, 1980, p. 219). B & F give simple conditions for handling (3.1b) and (3.1c) in the case that $\hat{F}_n$ is F_n. In the examples below I sketch how to verify (3.1) for the kernel estimator $\hat{F}_n$. Note that if $h_n \overset{wpl}{\to} 0$ and k is a density, then characteristic function arguments yield $\hat{F}_n \overset{wpl}{\Rightarrow} F$. If F is continuous, we can also conclude that $\|\hat{F}_n - F\|_\infty = \sup_x |\hat{F}_n(x) - F(x)| \overset{wpl}{\to} 0$.

EXAMPLE 1. The mean, $T(F) = \int x dF(x)$. In this case (3.1a) and (3.1b) are trivially satisfied and (3.1c) follows from (2.2) of B & F (or directly from the Lindeberg condition) if $\hat{F}_n \xrightarrow{wp1} F$ and $\int x^2 d\hat{F}_n \xrightarrow{wp1} \int x^2 dF(x) < \infty$. Since $\int x^2 d\hat{F}_n(x) = n^{-1}\Sigma X_i^2 + 2h_n\bar{X}\int xk(x)dx + h_n^2\int x^2k(x)dx$, it suffices to have $\int x^2k(x)dx < \infty$ and $h_n \xrightarrow{wp1} 0$.

EXAMPLE 2. Quantiles, $T(F) = F^{-1}(p) = \inf\{x: F(x) \geq p\}$, $0 < p < 1$. Here, $T(F;\Delta) = -\Delta(F^{-1}(p))/f(F^{-1}(p))$, and a modification of Ghosh's (1971) argument along with $\hat{f}_n \xrightarrow{wp1} f$ uniformly in a neighborhood of $F^{-1}(p)$ yields (3.1a). If h_n is not random, sufficient conditions for $\|\hat{f}_n - f\|_\infty \xrightarrow{wp1} 0$ are given in Silverman (1978), including f uniformly continuous and $h_n \to 0$, $(\log n)/nh_n \to 0$. Then (3.1b) is not necessary in this case since

$$T(\hat{F}_n;\hat{F}_n^* - \hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}[p - I(X_i^* \leq \hat{F}_n^{-1}(p))]/\hat{f}_n(\hat{F}_n^{-1}(p))$$

$$\stackrel{d}{=} \frac{1}{n}\sum_{i=1}^{n}[p - I(U_i \leq p)]/\hat{f}_n(\hat{F}_n^{-1}(p)) , \tag{3.2}$$

where the U_i are iid uniform (0,1) random variables. Of course (3.2) $\xrightarrow{d} N(0, p(1 - p)/f^2(F^{-1}(p))$ by the CLT and Slutsky's theorem.

EXAMPLE 3. Linear combinations of order statistics with smooth weight functions, $T(F) = \int_0^1 F^{-1}(t)J(t)dt$. Here, $T(F;\Delta) = -\int_{-\infty}^{\infty}\Delta(x)J(F(x))dx$. Restrict J to be bounded and continuous a.e. F^{-1} with at most a finite number of discontinuities. For T to be a translation statistic we need $\int_0^1 J(t)dt = 1$. If J vanishes in neighborhoods of 0 and 1, then (3.1a) and (3.1b) can be verified by slightly modifying the technique in Section 3 of Boos (1979) and (3.1c) follows from the triangular array CLT since $\hat{F}_n \xrightarrow{wp1} F$ implies

$$\iint[\hat{F}_n(x\wedge y) - \hat{F}_n(x)\hat{F}_n(y)]J(F(x))J(F(y))dxdy \tag{3.3}$$

$$\xrightarrow{wp1} \iint[F(x\wedge y) - F(x)F(y)]J(F(x))J(F(y))dx\,dy .$$

If J does not vanish in neighborhoods of 0 and 1, then $\int[\hat{F}_n(x)(1 - \hat{F}_n(x))]^{\frac{1}{2}}dx \xrightarrow{wp1} \int[F(x)(1 - F(x))]^{\frac{1}{2}}dx < \infty$ is sufficient to get (3.1a) and (3.1b). Since $\int x^2 d\hat{F}_n - (\int x d\hat{F}_n)^2 = \iint [\hat{F}_n(x \wedge y) - \hat{F}_n(x)\hat{F}_n(y)]dxdy$ and J is bounded, (3.3) and thus (3.1c) follow from the conditions in Example 1.

EXAMPLE 4. The Hodges-Lehmann functional, $T(F) = \text{median}\{F*F(2x)\}$, where $F*F$ denotes convolution. The usual estimator $T_n = \text{median}\{(X_i + X_j)/2,\ 1 \le i \le j \le n\}$ is asymptotically equivalent to $T(F_n)$. Here, $T(F;\Delta) = 2\int[F(2T(F) - x) - \frac{1}{2}]d\Delta(x)/G'(T(F))$, where $G'(x) = F*F(2x)$. As in Example 2, a modification of the Ghosh (1971) argument gives (3.1a) if $\|\hat{f}_n - f\|_\infty \overset{wp1}{\to}$ and $\|f\|_\infty < \infty$. (This case is slightly harder since it involves calculating the variance of certain U-statistics.) Then (3.1b) and (3.1c) follow under no further assumptions.

REMARK 3.2. At the beginning of this section, it was noted that one would typically use an approximation to the density $\hat{g}_n$ based on computer generated samples $X^* = (X_1^*,\cdots,X_n^*)$ from $\hat{F}_n$. Let T_i^*, $i = 1,\cdots,B$, be the values of $n^{\frac{1}{2}}[T(F_n^*) - T(\hat{F}_n)]$ calculated from B independent samples X^*. Then define the kernel estimator

$$(3.4) \qquad \hat{\hat{g}}_{n,B}(x) = \frac{1}{Bh_B}\sum_{i=1}^{B} k\left(\frac{x-T_i^*}{h_B}\right)$$

as the estimate of $\hat{g}_n$. It is not hard to believe that $\hat{\hat{g}}_{n,B}$ converges to g as n and $B \to \infty$ since Theorem 7 gives $\hat{g}_n \overset{wp1}{\to} g$ as $n \to \infty$. (However, a direct proof appears nontrivial.) But consider the following harder problem. Suppose that the X^* are generated from F_n so that $\hat{g}_n$ does not exist as a density wrt Lebesque measure since F_n is discrete. Can we hope for $\hat{\hat{g}}_{n,B}$ to still converge to g? The answer is *yes* under certain conditions on $\hat{G}_n$. Boos and Monahan (1982) observed this phenomenon in Monte Carlo work, but noted that the convergence in L_1 norm was slower when generating from F_n rather than from $\hat{F}_n$ in the case of the median and the Hodges-Lehmann estimator. Let us consider the conditional mean and variance of $\hat{\hat{g}}_{n,B}$.

$$E^*\hat{\hat{g}}_{n,B}(x) = \frac{1}{h_B}\int_{-\infty}^{\infty} k\left(\frac{x-y}{h_B}\right)d\hat{G}_n(y)$$

$$= \frac{1}{h_B}\int_{-\infty}^{\infty}\hat{G}_n(x - h_Bv)k'(v)dv$$

$$= \frac{1}{h_B}\int_{-\infty}^{\infty}\{\hat{G}_n(x-h_Bv) - G(x-h_Bv) - [\hat{G}_n(x) - G(x)]\}k'(v)dv$$

$$+ \int_{-\infty}^{\infty}\left\{\frac{G(x - h_Bv) - G(x)}{h_B}\right\}k'(v)dv .$$

We have used integration by part and $\int k'(v)dv = 0$. The last integral converges as $h_B \to 0$ to $-\int g(x)vk'(v)dv = g(x)$. The other integral suggests the condition

(3.5) $\{\hat{G}_n(x - h_Bv) - G(x - h_Bv) - [\hat{G}_n(x) - G(x)]\}/h_B \overset{wp1}{\to} 0$, each x and v,

as n and $B \to \infty$. One sufficient condition for (3.5) is $n^{\frac{1}{2}}\|\hat{G}_n - G\|_\infty = O(1)$, wp1, and $B^{\frac{1}{2}}h_B \to \infty$. For the variance we have

$$\mathrm{Var}^*\hat{\hat{g}}_{n,B}(x) \le \frac{1}{B}\mathrm{Var}^* \frac{1}{h_B} k\left(\frac{x-T_1^*}{h_B}\right)$$

$$\le \frac{1}{B}\left\{\frac{1}{h_B^2}\int_{-\infty}^{\infty}k^2\left(\frac{x-y}{h_B}\right)d\hat{G}_n(y) + [E^*\hat{\hat{g}}_{n,B}(x)]^2\right\}.$$

Thus, if k is bounded, $\int|vk'(v)|dv < \infty$, $h_B \to 0$ and $Bh_B^2 \to \infty$, then (3.5) gives $E^*[\hat{\hat{g}}_{n,B}(x) - g(x)]^2 \overset{wp1}{\to} 0$ and $\hat{\hat{g}}_{n,B}(x) \overset{P^*}{\to} g(x)$ (for almost all $X_1X_2,\ldots$). This proof holds for any $\hat{G}_n$ satisfying (3.5). A similar argument would work for the histogram estimators used by Efron (1979, p. 20).

4. Convergence of posterior densities.

The motivation for considering the convergence of densities of statistics arose in Boos and Monahan (1982). They considered the situation where one has observations from $f_0(x-\theta)$, f_0 is unknown but symmetric, $\hat{\theta}$ is a reasonable estimator of θ, and prior knowledge about θ is given by the prior density $\pi(\theta)$. Then an *estimated* Bayes posterior is obtained by using $\pi(\theta)$ and an estimated likelihood (essentially (3.4) evaluated at $\hat{\theta} = T_n(X)$). In this section I want to give a general formulation which justifies such an approach asymptotically.

Let $\theta \in \Theta$ be a subset of R^k. Suppose that there exists a sequence of estimators $\hat{\theta} = \hat{\theta}_n$ and constants $a_n \to \infty$ such that

(4.1) $$P(a_n(\hat{\theta} - \theta) \le x \mid \theta) \to L(x), \text{ each continuity point } x \text{ of } L.$$

Typically $a_n = n^{\frac{1}{2}}$ and L is a multivariate normal. Let L have density ℓ wrt Lebesque and let L_n denote the left-hand side of (4.1) or some estimate of that df with related density ℓ_n. Let $\pi(\theta)$ be a bounded and continuous prior density on R^k. Since $\hat{\theta}|\theta$ has (perhaps only approximately) density $a_n \ell_n(a_n(x - \theta))$, we can call $a_n \ell_n(a_n(\hat{\theta} - \theta))$ a likelihood and *define* the posterior density of $\theta|\hat{\theta}$ by

(4.2) $$p_n(t|\hat{\theta}) = \pi(t)a_n\ell_n(a_n(\hat{\theta} - t))/\int\pi(t)a_n\ell_n(a_n(\hat{\theta} - t))dt .$$

We are concerned about the limiting distribution of $a_n(\theta - \hat{\theta})|\hat{\theta}$ which has density $p_n(t/a_n + \hat{\theta}|\hat{\theta})$ and df

(4.3) $$P(a_n(\theta - \hat{\theta}) \le t|\hat{\theta}) = \int_{-t}^{\infty}\cdots\int\pi(\hat{\theta} - \frac{x}{a_n})dL_n(x)/\int_{R^d}\cdots\int\pi(\hat{\theta} - \frac{x}{a_n})dL_n(x).$$

The weakest type of result is that (4.3) converges weakly to 1 - L(-t). More interesting is the convergence of $p_n(t/a_n + \hat{\theta}|\hat{\theta})$ in topologies generated by pointwise convergence and $L_1^{(r)}$ convergence, $\int|t|^r|f - g|d\mu \to 0$. Bickel and

Yahav (1969) have considered convergence in $L_1^{(r)}$ for standard parametric Bayes posteriors.

Fix θ_0 as the value of θ in (4.1).

THEOREM 8. (i) If $L_n \Rightarrow L$ and $\hat{\theta} \overset{P}{\to} \theta_0$, then (4.3) $\overset{P}{\Rightarrow} 1 - L(-t)$. (ii) If $\ell_n \to \ell$ pointwise or in $L_1^{(r)}$ with $\int |t|^r \ell(t)dt < \infty$, then $p_n(t/a_n + \hat{\theta}|\hat{\theta}) \overset{P}{\to} \ell(-t)$ pointwise or in $L_1^{(r)}$. (iii) If $\hat{\theta} \overset{wp1}{\to} \theta_0$, then the convergences in (i) and (ii) can be strengthed to wp1.

PROOF. (i) By Skorohod construction we can find $\hat{\theta}^* \overset{d}{=} \hat{\theta}$ such that $\hat{\theta}^* \overset{wp1}{\to} \theta_0$. Fix $\omega \in \Omega_0 = \{\omega : \hat{\theta}^*(\omega) \to \theta_0\}$. Let $\theta_n = \hat{\theta}^*(\omega)$. First we consider the denominator of (4.3). On any compact subset A of R^k we can assert

$$(4.4) \qquad \int_A \pi(\theta_n - \frac{x}{a_n}) dL_n(x) \overset{wp1}{\to} \pi(\theta_0) \int_A dL(x)$$

since $\pi(\theta_n - x/a_n)$ is a bounded and continuous function which is converging uniformly to $\pi(\theta_0)$ on A. Then

$$\begin{aligned} \left|\int_{R^k} \pi(\theta_n - \frac{x}{a_n}) dL_n(x) - \pi(\theta_0)\right| &\le \left|\int_{R^k} [\pi(\theta_n - \frac{x}{a_n}) - \pi(\theta_n - \frac{x}{a_n}) I_A(x)] dL_n(x)\right| \\ &+ \left|\int_A \pi(\theta_n - \frac{x}{a_n}) dL_n(x) - \pi(\theta_0)\int_A dL(x)\right| \\ &+ \pi(\theta_0)(1 - \int_A dL(x)) . \end{aligned}$$

By choosing A and n suitably large, each of the above terms can be made arbitrarily small. Thus, the denominator of (4.3) $\overset{P}{\to} \pi(\theta_0)$. A similar argument yields that the numerator of (4.3) $\overset{P}{\to} \pi(\theta_0)(1 - L(-t))$. (ii) The convergences of ℓ_n all imply $L_n \Rightarrow L$ so that (i) takes care of the denominator of $p_n(t/a_n + \hat{\theta}|\hat{\theta})$. It remains to show that the numerator, $\pi(t/a_n + \hat{\theta})\ell_n(-t)$, converges to $\pi(\theta_0)$ in the appropriate senses. The pointwise convergence is obvious. For $L_1^{(r)}$

we have

$$\int_{R^k} |t|^r \pi(\frac{t}{a_n} + \theta_n)\ell_n(-t) - \pi(\theta_0)\ell(-t)|dt$$

$$\leq \int_{R^k} |t|^r|\pi(\frac{t}{a_n} + \theta_n) - \pi(\frac{t}{a_n}+\theta_n)I_A(t)|\ell_n(-t)dt$$

$$+ \int_A |t|^r|\pi(\frac{t}{a_n} + \theta_n) - \pi(\theta_0)|\ell_n(-t)dt$$

$$+ \pi(\theta_0) \int_A |t|^r|\ell_n(-t) - \ell(-t)|dt + \pi(\theta_0) \int_{R^k} |t|^r|1 - I_A(t)|\ell(-t)dt \quad ,$$

and each integral can be made suitably small by choosing A and n large. □

REMARKS 4.1. If L_n and ℓ_n are given by G_n and g_n or $\hat{G}_n$ and $\hat{g}_n$ of the forms in Sections 2 and 3, then Theorem 8 gives the asymptotic distribution of the associated Bayes posterior. From a practical standpoint it might be more interesting to let $\pi(\theta)$ grow more concentrated about some θ as n grows so that its affect is not negligible in large samples.

Acknowledgement. This paper was written during a sabbatical visit to the Department of Statistics, University California, Berkeley. I would like to thank Peter Bickel, Harold Goldstein, Kris Klaassen, and Lucien LeCam for helpful discussions during this time.

REFERENCES

Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā* Ser. A 29, 373-382.

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

Bickel, P. J. and Yahav, J. A. (1969). Some contributions to the asymptotic theory of Bayes solutions. *Z. Warsch. verw. Gebiete* 11, 257-276.

Boos, D. D. (1979). A differential of L-statistics. *Ann. Statist.* 7, 955-959.

Boos, D. D. and Monahan, J. F. (1982). A robust Bayesian approach using bootstrapped likelihoods. Preprint.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1-26.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* 42, 1957-1961.

Klaassen, C. A. J. (1982). Location estimators and spread. To appear in *Ann. Statist.*

Royden, H. L. (1968). *Real Analysis.* Macmillan, New York.

Silverman, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* 6, 177-184.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.