The Annals of Statistics
1974, Vol. 2, No. 6, 1308–1313

A LINEARIZED VERSION OF THE HODGES-LEHMANN ESTIMATOR[1]

BY ANDRÉ ANTILLE

University of California at Berkeley

In the estimation of a location parameter the Hodges–Lehmann estimator is known to have some "robust" properties, but it is very "expensive" for large sample sizes. By using the linearity of a special rank statistic we can find a linearized version which requires only $O(n \log n)$ operations.

0. Introduction. Let $X_1, X_2, \cdots$ be i.i.d. real random variables with density $f(x - \theta)$ where $f(x)$ is symmetrical about the origin. The Hodges–Lehmann estimator T has some "robust" properties (Bickel (1965)) but it is very "expensive" for large sample sizes: One needs at least $O(n^2)$ operations. The main purpose of this paper is to show that under some assumptions on f, there exists an estimator T_1 with the following properties:

(1) $n(T - T_1) = O_P(1)$, i.e., $n(T - T_1)$ is bounded in probability,
(2) T_1 requires only $O(n \log n)$ operations.

A new consistent estimator $\sigma_T{}^1$ of the asymptotic variance σ_T of T is given. Finally the speed of convergence of $\sigma_T{}^1(\sigma_T)^{-1}$ to 1 is investigated.

I thank Professor Bickel for pointing out to me that the paper of Kraft and van Eeden (1972) deals with similar topics. Their results are quite general, but for the case investigated in this paper we obtain much more precise information about the behavior of the linearized version.

1. Preliminaries. The whole work relies on the asymptotic linearity of a special rank statistic (Antille (1972)):

Let $X_1, X_2, \cdots$ be i.i.d. real random variables with symmetrical density $f(x)$.

Define: $h(\bar{x}) = n^{-\frac{3}{2}} \sum [I(x_i + x_j \leqq 0) - 2^{-1}]$, where the summation extends over all $i < j$ with $1 \leqq i, j \leqq n$.

$$S_n(\bar{x}, t) = h(\bar{x} - tn^{-\frac{1}{2}}) - h(\bar{x}) .$$

$I(A)$ means the indicator function of A and $\bar{x}$ the vector $(x_1, x_2, \cdots, x_n)$. We should write $h_n(\bar{x})$ or $h(\bar{x}_n)$, but we drop the index n in order to simplify the notation.

Received June 1973; revised November 1973.

[1] This paper was supported by the Swiss National Science Foundation. This research was supported in part by the Office of Naval Research, Contract NONR N00014-69-A-0200-1038.

AMS 1970 *subject classifications.* Primary 62G05, 62G25; Secondary 62E15, 60F05.

Key words and phrases. One-sample problem, linear rank statistic, asymptotic variance, weak convergence of stochastic processes.

Consider now the process

$$Y_n(t) = n^{\frac{1}{2}}[S_n(\bar{x}, t) - t \int f^2(x)\, dx] ,$$

for $t \in [-M, +M]$ where M is an arbitrary fixed number.

We have then the following theorem:

If we assume

(i) $$\int f^3(x)\, dx < \infty ,$$

(ii) $$\Delta^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta} [f(y + z) - f(y)]^2\, dz\, dy \to_{\Delta \to 0} 0 ,$$

then the process $(Y_n(t))_{t \in [-M, +M]}$ converges weakly to a process of the form $(tZ)_{t \in [-M, +M]}$ where Z is a real random variable with normal distribution $N(0, c^2)$ where

$$c^2 = 4[\int f^3(x)\, dx - (\int f^2(x)\, dx)^2] .$$

REMARK. Note that condition (ii) is satisfied if, for example,

(a) f is such that $|f(x + t) - f(x)| \leqq |t|^{\alpha} g(x)$, with $\alpha > 2^{-1}$ and $g(x) \in L_2(-\infty, +\infty)$ or

(b) f is absolutely continuous and $f' \in L_2(-\infty, +\infty)$.

2. The estimator T_1. Let $X_1, X_2, \cdots$ be i.i.d. random variables with density $f(x - \theta)$ where $f(x)$ is symmetrical about the origin.

Then the Hodges–Lehmann estimator T is essentially the value of t for which

$$h(\bar{x} - t) = 0 .$$

Now let t be fixed. One needs $O(n \log n)$ operations to order the sample $|X_1 - tn^{-\frac{1}{2}}|, \cdots, |X_n - tn^{-\frac{1}{2}}|$.

Hence $h(\bar{x} - t)$, for a fixed t, can be calculated in $O(n \log n)$ operations. This follows from the relation:

$$\sum R_i^+ = \sum I(X_i + X_j > 0) ,$$

where the summation on the left-hand side extends over all $1 \leqq i \leqq n$ such that $X_i > 0$, and the summation on the right-hand side over all $i \leqq j$ with $1 \leqq i, j \leqq n$. R_i^+ means the rank of $|X_i|$ in the sample $|X_1|, |X_2|, \cdots, |X_n|$.

Now let T_0 be some invariant estimator for θ with the property that $n^{\frac{1}{2}}(T_0 - \theta)$ converges weakly to some distribution. (For example, take for T_0 the median of $X_1, X_2, \cdots, X_n$.) Then, if $S_n(\bar{x} - T_0, 1) \neq 0$, define T_1 as the solution of the following equation:

$$h(\bar{x} - T_0) + n^{\frac{1}{2}}(t - T_0) S_n(\bar{x} - T_0, 1) = 0 ;$$

otherwise set $T_1 = T_0$.

We have the following:

THEOREM 1. *If we assume property* (i) *stated in Section* 1 *and*

(ii)′ $$\limsup_{\Delta \downarrow 0} \Delta^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta} [f(y + z) - f(y)]^2\, dz\, dy < \infty ,$$

then $n(T - T_1) = O_P(1)$.

We first prove the following:

LEMMA 1. *If we assume properties* (i) *and* (ii)′ *stated in Theorem* 1, *then*

$$\sup_{|t|\leq M} |Y_n(t)| = O_P(1) ,$$

for every fixed number M.

PROOF. Let $Z_n(t) = Y_n(t) - E[Y_n(t)]$. Then under assumption (i) we have (Antille (1972)):

(A) The process $(Z_n(t))_{t\in[-M,+M]}$ converges weakly to the process $(tZ)_{t\in[-M,+M]}$,

(B) $$\sup_{|t|\leq M} |EY_n(t)| \leqq n^{-\frac{1}{2}} M \int f^2(x)\,dx + M^2\Delta_n^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta_n} [f(x+z) - f(x)]^2\,dz\,dx ,$$

with $\Delta_n = 2n^{-\frac{1}{2}}M$.

Now by using assumption (ii)′ we get:

$$\sup_{|t|\leq M} |Y_n(t)| \leqq \sup_{|t|\leq M} |Z_n(t)| + M^2K ,$$

for some finite positive constant K.

It is now easy to prove Theorem 1. We can assume that the true value of θ is 0. We have:

$$h(\bar{x} - T) = 0$$

and this is the same as

$$S_n(\bar{x}, n^{\frac{1}{2}}T) + h(\bar{x}) = 0 , \tag{1}$$

or

$$n^{-\frac{1}{2}}Y_n(n^{\frac{1}{2}}T) + n^{\frac{1}{2}}T \cdot \int f^2(x)\,dx + h(\bar{x}) = 0 . \tag{2}$$

By definition T_1 satisfies asymptotically the following equation:

$$n(T_1 - T_0)S_n(\bar{x} - T_0, 1) + n^{\frac{1}{2}}h(\bar{x} - T_0) = 0 . \tag{3}$$

By subtracting (2) from (3) we get:

$$\begin{aligned} n(T - T_1)S_n(\bar{x} - T_0, 1) &= -Y_n(n^{\frac{1}{2}}T) - nT\left[\int f^2(x)\,dx - S_n(\bar{x} - T_0, 1)\right] \\ &\quad + n^{\frac{1}{2}}[h(\bar{x}) - h(\bar{x} - T_0)] + nT_0 \cdot S_n(\bar{x} - T_0, 1) , \end{aligned} \tag{4}$$

which is the same as:

$$\begin{aligned} n(T - T_1)S_n(\bar{x} - T_0, 1) &= -Y_n(n^{\frac{1}{2}}T) - Y_n(n^{\frac{1}{2}}T_0) + [n^{\frac{1}{2}}T_0 + n^{\frac{1}{2}}T][Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0)] . \end{aligned} \tag{5}$$

The statistic $S_n(\bar{x} - T_0, 1)$ which appears in the left-hand side of (5) converges to $\int f^2(x)\,dx$ in probability. This follows from the relation

$$|S_n(\bar{x} - T_0, 1) - \int f^2(x)\,dx| = n^{-\frac{1}{2}}|Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0)| ,$$

using Lemma 1 and the fact that the family $\{n^{\frac{1}{2}}T_0\}_{n=1,2,\ldots}$ is tight.

It is therefore sufficient to show that the right-hand side of (5) is an $O_P(1)$.

But this is a direct consequence of Lemma 1, using once more the fact that the families $\{n^{\frac{1}{2}}T_0\}_{n=1,2,\ldots}$ and $\{n^{\frac{1}{2}}T\}_{n=1,2,\ldots}$ are tight.

3. The estimator $\sigma_T{}^1$. Let $\sigma_T{}^0$ be the Lehmann estimator for σ_T (Lehmann (1963)). Define

$$\sigma_T{}^1 = (12)^{-\frac{1}{2}}[S_n(\bar{x} - T_0, 1)]^{-1} \qquad \text{if} \quad S_n(\bar{x} - T_0, 1) \neq 0 ,$$
$$\sigma_T{}^1 = \sigma_T{}^0 \qquad \text{otherwise.}$$

We have the following:

THEOREM 2. *Under the same assumptions as in Theorem* 1 *we have*:

$$n^{\frac{1}{2}}[\sigma_T{}^1(\sigma_T)^{-1} - 1] = O_P(1) .$$

PROOF. Assume again that the true value of θ is 0. We have:

$$n^{\frac{1}{2}}[\sigma_T{}^1(\sigma_T)^{-1} - 1] \cdot S_n(\bar{x} - T_0, 1) = n^{\frac{1}{2}}[\int f^2(x)\,dx - S_n(\bar{x} - T_0, 1)] .$$

By definition of $Y_n(t)$,

$$n^{\frac{1}{2}}[\int f^2(x)\,dx - S_n(\bar{x} - T_0, 1)] = Y_n(n^{\frac{1}{2}}T_0) - Y_n(1 + n^{\frac{1}{2}}T_0) .$$

Since we already proved in Theorem 1 that $S_n(\bar{x} - T_0, 1)$ converges to $\int f^2(x)\,dx$ in probability, and that $Y_n(n^{\frac{1}{2}}T_0) - Y_n(1 + n^{\frac{1}{2}}T_0) = O_P(1)$, the proof is complete.

Finally we have the following:

THEOREM 3. *If f has the properties* (i) *and* (ii) *stated in Section* 1, *then*

$$n^{\frac{1}{2}}[\sigma_T{}^1(\sigma_T)^{-1} - 1]$$

has asymptotically a normal distribution $N(0, e^2)$, where $e^2 = c^2[\int f^2(x)\,dx]^{-2}$.

PROOF. Assume again that the true value of θ is 0. Since assumption (ii)′ is weaker than (ii), $S_n(\bar{x} - T_0, 1)$ converges in probability to $\int f^2(x)\,dx$. It is therefore sufficient to show that $Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0)$ has asymptotically a normal distribution $N(0, c^2)$.

Let $\varepsilon > 0$ and x be arbitrary numbers. The family $\{n^{\frac{1}{2}}T_0\}_{n=1,2,\ldots}$ is supposed to be tight. Hence there exists an M (M depends on ε) such that

$$P\{1 + |n^{\frac{1}{2}}T_0| \leqq M\} \geqq 1 - \varepsilon .$$

We have then:

$$\begin{aligned} P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x\} & \\ \leqq P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x,\ 1 + |n^{\frac{1}{2}}T_0| \leqq M\} + \varepsilon & \\ \leqq P\{\min_{t\in[-M,+M]}[Y_n(1 + t) - Y_n(t)] < x\} + \varepsilon . & \end{aligned}$$

If n goes to infinity, we get by using the weak convergence of $Y_n(t)$ to tZ:

$$\lim_n \sup P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x\} \leqq P\{Z < x\} + \varepsilon .$$

By using the inequality

$$\begin{aligned} P\{\max_{t\in[-M,+M]}[Y_n(1 + t) - Y_n(t)] < x\} & \\ \leqq P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x\} + \varepsilon , & \end{aligned}$$

we obtain in the same way as before:

$$P\{Z < x\} \leqq \lim_n \inf P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x\} + \varepsilon .$$

ε is arbitrary. Hence

$$\lim_{n\to\infty} P\{Y_n(1 + n^{\frac{1}{2}}T_0) - Y_n(n^{\frac{1}{2}}T_0) < x\} = P\{Z < x\} ,$$

for every x and the proof is complete.

4. Remarks.

(1) Assumption (i) does not imply (ii). For example, if we take

$$f(x) = 1 \qquad \text{for} \quad |x| \leqq 2^{-1} ,$$
$$f(x) = 0 \qquad \text{for} \quad |x| > 2^{-1} ,$$

we get by easy calculation:

$$\Delta^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta} [f(x + y) - f(x)]^2 \, dy \, dx = 1 \qquad \text{for all} \quad \Delta > 0 .$$

(2) There exist densities which satisfy (i) but not (ii)′. For example, take:

$$f(x) = d|x|^{-\frac{1}{6}} \qquad \text{for} \quad |x| \leqq 1 ,$$
$$f(x) = 0 \qquad \text{for} \quad |x| > 1 .$$

(d is a constant such that $\int f(x)\, dx = 1$.)

For Δ small enough we have then:

$$\begin{aligned} \Delta^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta} [f(x + y) - f(x)]^2 \, dy \, dx &\geqq d^2\Delta^{-2} \int_{\Delta^2-1}^{\Delta} \int_0^{\Delta} [x^{-\frac{1}{6}} - (x + y)^{-\frac{1}{6}}]^2 \, dy \, dx \\ &\geqq d^2\Delta^{-2} \int_{\Delta^2-1}^{\Delta} \int_0^{\Delta} [6^{-1}(\tfrac{3}{2})^{-\frac{7}{6}} y \Delta^{-\frac{7}{6}}]^2 \, dy \, dx \\ &= a(\Delta^{-\frac{1}{3}}) \end{aligned}$$

where a is a constant.

(3) Condition (ii)′ is satisfied if for example f has the following property:

There exists a finite set of points $\varepsilon(-\infty, +\infty)$, say $t_1 < t_2 < \cdots < t_k$, such that

1. For every j, $1 \leqq j \leqq k$, $\lim_{\delta \downarrow 0} f(t_j + \delta)$ and $\lim_{\delta \uparrow 0} f(t_j - \delta)$ exist and are finite.

2. In every interval $(-\infty, t_1), (t_1, t_2), \cdots, (t_k, \infty)$ f satisfies condition (a) (with $\alpha \geqq 2^{-1}$) or (b) given in the remark at the end of Section 1.

PROOF. It is sufficient to prove for the case where $k = 1$. Assume for example that condition (b) is satisfied in the intervals $(-\infty, t_1)$, (t_1, ∞).

Let $g(x, y) = [f(x + y) - f(x)]^2$ and $t = t_1$. We have then:

$$\begin{aligned} \Delta^{-2} \int_{-\infty}^{+\infty} \int_0^{\Delta} g(x, y) \, dy \, dx &= \Delta^{-2} \int_{-\infty}^{t-2\Delta} \int_0^{\Delta} g(x, y) \, dy \, dx + \Delta^{-2} \int_{t-2\Delta}^{t+2\Delta} \int_0^{\Delta} g(x, y) \, dy \, dx \\ &\quad + \Delta^{-2} \int_{t+2\Delta}^{\infty} \int_0^{\Delta} g(x, y) \, dy \, dx \\ &= A + B + C . \end{aligned}$$

Now $A = \Delta^{-2} \int_{-\infty}^{t-2\Delta} \int_0^{\Delta} [\int_0^y f'(z + x)\, dz]^2 \, dy \, dx$.

By using the Schwarz inequality and the theorem of Fubini we obtain:

$$A \leqq \int_{-\infty}^{t} f'^2(u)\, du \cdot 3^{-1}\Delta .$$

In the same way we get a similar result for C. By assumption 1 we obtain, for Δ small enough:

$$B \leqq K\Delta^{-2} \int_{t-2\Delta}^{t+2\Delta} \int_{0}^{\Delta} dy\, dx = 4K$$

where K is some finite constant.

REFERENCES

[1] ANTILLE, A. (1972). Linéarité asymptotique d'une statistique de rang. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiète* **24** 309–324.

[2] BICKEL, P. J. (1965). On some robust estimates of location. *Ann. Math. Statist.* **36** 847–858.

[3] KRAFT, C. H. and VAN EEDEN, C. (1972). Linearized rank estimates and signed-rank estimates for the general linear hypothesis. *Ann. Math. Statist.* **43** 42–57.

[4] LEHMANN, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34** 1507–1512.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720