ON THE HODGES AND LEHMANN SHIFT ESTIMATOR IN THE TWO SAMPLE PROBLEM[1]

BY TERRENCE FINE[2]

University of California, Berkeley

1. Introduction. This note provides a characterization of the Hodges and Lehmann (1960) estimator of shift in the two sample problem, Δ^*, and suggests an alternative estimator Δ_ϵ. The asymptotic variance of Δ_ϵ is never much more than that of Δ^* and for some underlying distributions it can be indefinitely smaller (Theorem 3).

It is assumed that we are given a sample $X_1, \cdots, X_n$ of observations that are iid as $F(x)$ and a second sample, independent of the first, of observations $Y_1, \cdots, Y_m$ that are iid as $G(x)$. Furthermore, it is assumed that for some initially unknown shift Δ, $F(x - \Delta) = G(x) \varepsilon \mathcal{G}$, where $\mathcal{G}$ is the class of all absolutely continuous distributions, and $\mathcal{G}$ is otherwise unspecified. Let $N = n + m$ and $\lambda_N = n/N$. The final assumption is that for some λ_0, $0 < \lambda_0 \leqq \lambda_N \leqq 1 - \lambda_0 < 1$. It is then desired to estimate Δ.

The empirical distributions for the samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ will be respectively denoted by $F_n(x)$ and $G_m(x)$. The total sample empirical distribution $H_N(x)$ when the sample $X_1, \cdots, X_n$ is shifted to the right by an amount $\bar{\Delta}$, is given by

$$H_N(x) = \lambda_N F_n(x - \bar{\Delta}) + (1 - \lambda_N) G_m(x),$$

where in our notation the dependence of H_N upon $\bar{\Delta}$ is implicit.

2. A characterization of Δ^*. The shift estimator $\Delta^* = \text{med}\,(Y_i - X_j)$ $(i = 1, \cdots, m; j = 1, \cdots, n)$ has been proposed and examined by Hodges and Lehmann (1960). The asymptotic distribution of Δ^* is normal with mean Δ and asymptotic variance $\sigma^2_{\Delta^*}$ given by

$$\sigma^2_{\Delta^*} = (12\lambda_N(1 - \lambda_N)N)^{-1}[\textstyle\int_{-\infty}^{\infty} G'\, dG]^{-2}.$$

A characterization of Δ^* is provided by

THEOREM 1. *The estimator Δ^* minimizes the two-sample version of the Cramér-von Mises statistic*

$$W_N^{\,2} = \textstyle\int_{-\infty}^{\infty} [F_n(x - \bar{\Delta}) - G_m(x)]^2\, dx. \tag{1}$$

PROOF. Choose some $A > \max\,(X_1 + \bar{\Delta}, \cdots, X_n + \bar{\Delta}, Y_1, \cdots, Y_m)$ and replace the empirical distributions in (1) by their unit step definitions.

Equation (1) becomes

$$W_N^{\,2} = \textstyle\int_{-\infty}^{A} n^{-2} \sum \sum_{i,j=1}^{n} U(x - X_i - \bar{\Delta}) U(x - X_j - \bar{\Delta})\, dx$$

Received 18 October 1965; revised 14 May 1966.

[1] This work was supported by the Adolph C. and Mary S. Miller Institute, University of California, Berkeley.

[2] Now at Cornell University.

$$(2) \qquad + \int_{-\infty}^{A} m^{-2} \sum \sum_{i,j=1}^{m} U(x - Y_i) U(x - Y_j)\, dx$$
$$- 2(nm)^{-1} \int_{-\infty}^{A} \sum_{i=1}^{n} \sum_{j=1}^{m} U(x - X_i - \bar{\Delta}) U(x - Y_j)\, dx.$$

Evaluation of (2) after reduction yields

$$W_N^2 = -n^{-2} \sum \sum_{i,j=1}^{n} \max(X_i, X_j) - m^{-2} \sum \sum_{i,j=1}^{m} \max(Y_i, Y_j)$$
$$+ (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (X_i + Y_j) + (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} |Y_j - X_i - \bar{\Delta}|.$$

The only term involving $\bar{\Delta}$ is $(nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} |Y_j - X_i - \bar{\Delta}|$, and it's well known that this term is minimized by $\bar{\Delta} = \operatorname{med}(Y_j - X_i)$, as claimed.

3. An alternative shift estimator. A shift estimator Δ_ϵ, having desirable properties when compared to Δ^*, can be generated as follows. Define the statistic

$$(3) \qquad V_N(\bar{\Delta}) = \int_{0<H_N<1} J(H_N)\, d(F_n(x - \bar{\Delta}) - G_m(x)),$$

where $J(x)$ is such that for some $0 < \epsilon < \frac{1}{2}$

$$(4) \qquad \begin{aligned} J^{(2)}(x) &= x^{-2} && \text{if } 0 \leqq x < \epsilon \\ &= 0 && \text{if } \epsilon \leqq x < 1 - \epsilon \\ &= -(1 - x)^{-2} && \text{if } 1 - \epsilon \leqq x \leqq 1 \end{aligned}$$

and $J^{(1)}(\epsilon) = -\epsilon^{-1}$; V_N is closely related to T_N of Chernoff and Savage. Following the technique of Hodges and Lehmann (1963) for the conversion of a test statistic to a point estimator, we define

$$(5) \qquad \Delta_\epsilon = \sup \{\bar{\Delta}: V_N(\bar{\Delta}) > 0\}.$$

The definition of Δ_ϵ in (5) is partially justified by

LEMMA 1. *If $J(x)$ is non-increasing in x, then $V_N(\bar{\Delta})$ is non-increasing in $\bar{\Delta}$.*

PROOF. $Z_1, \cdots, Z_N$ is the ordered sample of elements $X_1 + \bar{\Delta}, \cdots, X_n + \bar{\Delta}$, $Y_1, \cdots, Y_m$, and we define

$$\begin{aligned} Z_{Ni} &= 1 && \text{if } Z_i \text{ is an } X + \bar{\Delta} \\ &= 0 && \text{if } Z_i \text{ is a } Y. \end{aligned}$$

Then $V_N(\bar{\Delta})$ can be expressed as

$$\begin{aligned} V_N(\bar{\Delta}) = {} & N^{-1} J((N - 1)/N)[(1 - \lambda_N)^{-1} - Z_{NN}/\lambda_N(1 - \lambda_N)] \\ & + N^{-1} \sum_{i=2}^{N-1} [J((i - 1)/N) - J(i/N)] \\ & \cdot \{[\lambda_N(1 - \lambda_N)]^{-1} \sum_{j=1}^{i} Z_{Nj} - i/(1 - \lambda_N)\}. \end{aligned}$$

As $\bar{\Delta}$ increases, the number of $X + \bar{\Delta}$ terms in the first i terms of the ordered total sample is non-increasing. Hence, $\sum_{j=1}^{i} Z_{Nj}$ is non-increasing. By hypothesis J is non-increasing and thus $J((i - 1)/N) - J(i/N)$ is non-negative. Since V_N is a sum of non-negatively weighted, non-increasing terms, it is non-increasing in $\bar{\Delta}$ as claimed.

The asymptotic behavior of Δ_ϵ follows from that of V_N and that of V_N is given by

LEMMA 2. *The statistic* $V_N(\bar{\Delta})$ *is asymptotically normally distributed with variance* $\sigma_V^2 = O(N^{-1})$ and mean

$$(6) \qquad m_V(\bar{\Delta}) = \int_{0<H<1} [G(x) - F(x - \bar{\Delta})]J'(H(x))\, dH(x),$$

where $H(x) = \lambda_N F(x - \bar{\Delta}) + (1 - \lambda_N)G(x)$.

PROOF. Theorem 1 of Chernoff and Savage and the details of its proof establish the asymptotic equivalence

$$V_N = \int_{0<H<1} (H_N - H)J'(H)\, d(F - G) - \int_{0<H<1} (F - G)J'(H)\, dH \\ - \int_{0<H<1} (F_n - F + G - G_m)J'(H)\, dH + o_p(N^{-\frac{1}{2}}).$$

The mean and variance follow by direct calculation, although we omit the cumbersome expression for σ_V^2. Asymptotic normality follows essentially from the asymptotic normality of $N^{\frac{1}{2}}(F_n - F)$ and $N^{\frac{1}{2}}(G_m - G)$, or reference to Chernoff and Savage.

A sufficient condition for the asymptotic normality of Δ_ϵ is

THEOREM 2. *If*

$$(7) \qquad \lim_{N\to\infty} N^{\frac{1}{2}} m_V(\Delta + \alpha N^{-\frac{1}{2}}) = \alpha \int G' J'(G)\, dG,$$

then Δ_ϵ *is asymptotically normally distributed with asymptotic mean* Δ *and asymptotic variance*

$$\sigma_\epsilon^2 = [2/\lambda_N(1 - \lambda_N)N][\int_{-\infty}^{\infty} G'J'(G)\, dG]^{-2} \iint_{0\leqq x<y\leqq 1} x(1 - y)J'(x)J'(y)\, dx\, dy.$$

PROOF. From (5) and the fact that V_N is non-increasing, it's immediate that the event $\Delta_\epsilon \leqq \Delta + \alpha N^{-\frac{1}{2}}$ is equivalent to the event $V_N(\Delta + \alpha N^{-\frac{1}{2}}) > 0$. By the asymptotic normality of V_N this establishes that

$$P[\Delta_\epsilon - \Delta \leqq \alpha N^{-\frac{1}{2}}] = \text{cerf } [-(\alpha N^{-\frac{1}{2}})(\int_{-\infty}^{\infty} G'J'(G)\, dG)/\sigma_V],$$

and the theorem follows after some calculation and approximation to σ_V. For our choice of $J(x)$, and assuming the validity of (7), we have that

$$(8) \quad \sigma_\epsilon^2 = [12\lambda_N(1 - \lambda_N)N\epsilon^2]^{-1}[1 + 12\epsilon^2 + 16\epsilon^3][\int_{-\infty}^{\infty} G'J'(G)\, dG]^{-2}.$$

In order to compare the performance of Δ_ϵ and Δ^* in terms of their asymptotic variances we establish

THEOREM 3. *Under the above definitions and assumptions* (*less* (7))

$$(9) \qquad \sup_G [\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2)] = \infty,$$

and

$$(10) \qquad \inf_G [\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2)] = 1 - O(\epsilon^2).$$

PROOF. Introduce the sequence of absolutely continuous distributions $\{E_n\}$ with densities defined by

$$
\begin{aligned}
E_n'(x) &= 2\delta_n X && \text{if } 0 \leqq x \leqq (t/\delta_n)^{\frac{1}{2}} \\
&= \delta_n^{-\frac{1}{2}} && \text{if } (t/\delta_n)^{\frac{1}{2}} < x \leqq (t/\delta_n)^{\frac{1}{2}} + \delta_n \\
&= 2\delta_n(A - x) && \text{if } (t/\delta_n)^{\frac{1}{2}} + \delta_n < x \leqq A = \delta_n + (t/\delta_n)^{\frac{1}{2}} \\
& && \quad + [(1 - t - \delta_n^{\frac{1}{2}})/\delta_n]^{\frac{1}{2}} \\
&= 0 && \text{if } x \text{ otherwise,}
\end{aligned} \tag{11}
$$

where δ_n is any sequence decreasing to zero. This sequence has the property, as can be verified by integrating and taking limits, that for $0 < t < 1$,

$$
\begin{aligned}
\lim_{n\to\infty} [\textstyle\int_{-\infty}^{\infty} & E_n'J'(E_n)\, dE_n / \textstyle\int_{-\infty}^{\infty} E_n'\, dE_n] \\
&= \lim_{n\to\infty} \{[J'(t) + O(\delta_n^{\frac{1}{2}})]/([1 + O(\delta_n^{\frac{1}{2}})]\} = J'(t).
\end{aligned}
$$

By properly selecting t, the limit of this sequence will yield the asserted sup and inf.

To verify (9) observe that when (7) holds

$$
\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2) = \epsilon^2[1 + 12\epsilon + 16\epsilon^3]^{-1}\{\textstyle\int_{-\infty}^{\infty} G'J'(G)\, dG/\textstyle\int_{-\infty}^{\infty} G'\, dG\}^2. \tag{12}
$$

Since (7) is valid for any of the E_n, as is verifiable by integration in (6), let us consider the sequence K_n defined as E_n but with $t = \delta_n$. For each n (12) is valid, and as $n \to \infty$,

$$
\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2) \propto [J'(\delta_n)]^2 \to [J'(0)]^2 = \infty.
$$

This verifies the supremum result.

The proof of the infimum condition is similar and rests on the observation that for $G' \geqq 0$

$$
(\textstyle\int_{-\infty}^{\infty} G'J'(G)\, dG)^2 \geqq \min_t [J'(t)]^2(\textstyle\int_{-\infty}^{\infty} G'\, dG)^2 = \epsilon^{-2}(\textstyle\int_{-\infty}^{\infty} G'\, dG)^2.
$$

Thus a lower bound to $\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2)$ is $[1 + 12\epsilon^2 + 16\epsilon^3]^{-1}$. To prove that this lower bound is the infimum we need only exhibit a sequence of distributions such that $\lim_{N\to\infty} (\sigma_{\Delta^*}^2/\sigma_\epsilon^2)$ converges to this lower bound. The sequence $\{E_n\}$ with $t = \frac{1}{2}$ will obviously suffice to complete the proof.

If G is a uniform distribution, then one can show that $\sigma_\epsilon^2 = o(N^{-1})$ and Δ_ϵ is asymptotically non-normal. Hence the supremum is attainable within the class of allowed distributions.

Acknowledgment. The author is indebted to Professor E. Lehmann of the University of California, Berkeley, Statistics Department, for introducing him to this problem.

REFERENCES

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29** 972–994.

Hodges, J. L., Jr., and Lehmann, E. L. (1960). Comparison of the normal scores and Wilcoxon tests. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 307–317. Univ. of California Press.

Hodges, J. L., Jr., and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.