NONPARAMETRIC CONFIDENCE INTERVALS FOR A SHIFT PARAMETER[1]

By E. L. Lehmann

University of California, Berkeley

Summary. Exact expressions and large-sample approximations are given for the nonparametric confidence intervals for a shift parameter Δ, which are obtained from the two-sample Wilcoxon test. These intervals are shown to have the same asymptotic efficiency relative to the standard confidence intervals for Δ as the Wilcoxon test has relative to Student's t-test. As a consequence of this result, a constant multiple of the length of the nonparametric intervals is shown to be a consistent estimator of the quantity $1/\int f^2(x)\,dx$.

Let $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ be independent observations from distributions $P[X_i \leq x] = F(x)$ and $P[Y_j \leq y] = F(y - \Delta)$ respectively, where F is assumed to be continuous but otherwise unknown. Exact confidence intervals for Δ can be based on nonparametric tests such as the Wilcoxon test. Let $U(\Delta)$ be the number of differences $Y_j - X_i$ which exceed Δ, and suppose that the two-sided symmetric level α test of the hypothesis $H(\Delta_0) : \Delta = \Delta_0$ accepts $H(\Delta_0)$ when

$$(1) \qquad C_\alpha \leq U(\Delta_0) \leq mn - C_\alpha.$$

If the ordered set of mn differences $Y_j - X_i$ is denoted by

$$D^{(1)} < \cdots < D^{(mn)},$$

the confidence intervals obtained from (1) by solving for Δ_0 and replacing Δ_0 by Δ are seen to be

$$(2) \qquad \Delta_L < \Delta < \Delta_U$$

where

$$(3) \qquad \Delta_L = D^{(C_\alpha)} \quad \text{and} \quad \Delta_U = D^{(mn+1-C_\alpha)}.$$

For given sample sizes m and n, a constant C_α for which (1) has exactly probability $1 - \alpha$ may not exist. For the large-sample problems to be considered below it is enough that the constants $C_\alpha = C_\alpha(m, n)$ are chosen in such a way that the probability $1 - \alpha(m, n)$ of (1) tends to the specified value $1 - \alpha$ as m and n tend to infinity.

For small and moderate values of m and n, the critical value C_α can be read from tables of the null distribution of the Wilcoxon statistic. For large m and n, one has the following approximation.

LEMMA 1. *The critical value C_α is for large m and n equal to*

$$(4) \qquad C_\alpha = mn/2 - K_{\alpha/2}[mn(m+n)/12]^{\frac{1}{2}} + o[mn(m+n)]^{\frac{1}{2}}$$

where $K_{\alpha/2}$ is the $100\alpha/2$ upper percentage point of the standard normal distribution.

PROOF. Since $\Delta_L = \sup\{\Delta : U(\Delta) > mn - C_\alpha\}$ we have $\alpha/2 = \lim P_\Delta[\Delta < \Delta_L] = \lim P_\Delta[U(\Delta) > mn - C_\alpha]$ and the result now follows from the fact that the Wilcoxon statistic $U(0)$ for $\Delta = 0$ has a normal limiting distribution with mean $mn/2$ and variance $mn(m+n+1)/12$.

What is the efficiency of the intervals (2) relative to the standard intervals

$$(5) \quad \bar{Y} - \bar{X} - K_{\alpha/2}(1/m + 1/n)^{\frac{1}{2}}S < \Delta < \bar{Y} - \bar{X} + K_{\alpha/2}(1/m + 1/n)^{\frac{1}{2}}S$$

where S^2 is the usual estimate of the variance σ^2 of F? Suppose first that efficiency is measured in terms of the probability of covering false values, say in terms of the probability that the intervals cover the value $\Delta + rN^{-\frac{1}{2}}$, where $N = m + n$. Then it follows from the correspondence between confidence intervals and the tests on which they are based, and the known efficiency properties of the Wilcoxon test, that if the intervals defined by (2) and (3) are based on $m = \rho N$ and $n = (1 - \rho)N$ observations, and the intervals (5) on $m' = \rho N'$ and $n' = (1 - \rho)N'$ observations, they will have the same limiting probability of covering the values $\Delta + rN^{-\frac{1}{2}}$, provided

$$(6) \qquad N'/N \to 12\sigma^2 \left(\int f^2(x)\,dx \right)^2$$

as $N \to \infty$, where f is the density of the distribution F. In this sense, (6) is the asymptotic relative efficiency of the two sets of intervals.

Alternatively, efficiency might be measured in terms of the lengths of the intervals. To carry out this comparison, we shall now prove the following result.

THEOREM 1. *If $L = \Delta_U - \Delta_L$ is the length of the intervals defined by (2) and (3) and based on $m = \rho N$ and $n = (1 - \rho)N$ observations, then*

$$(7) \qquad N^{\frac{1}{2}}L \to K_{\alpha/2}/[3\rho(1-\rho)]^{\frac{1}{2}} \int f^2(x)\,dx$$

in probability as $N \to \infty$.

To prove this result, we shall require the following lemmas.

LEMMA 2. *Let $T_{1N}, \cdots, T_{NN}(N = 1, 2, \cdots)$ be a sequence of sets of random variables (not necessarily identically or independently distributed). Suppose for any fixed a and b the number of T's between a/N^β and b/N^β equals $[k(b-a) + R_N]N^\gamma$ where R_N tends to zero in probability uniformly for all a, b satisfying $-A < a < b < A$ for any finite A. Let U_N be a sequence of random variables (possibly depending on $T_{N1}, \cdots, T_{NN}$) which are bounded in probability, and let Z_N be a value such that exactly $cN^\gamma + o(N^\gamma)$ of the T's lie between U_N/N^β and Z_N/N^β. Then $Z_N - U_N \to c/k$ in probability.*

PROOF. Suppose to the contrary that with probability exceeding ϵ, the inequality $Z_N - U_N > (c + \delta)/k$ holds for some positive δ and ϵ for arbitrarily

large N. (The proof for the case $Z_N - U_N < (c - \delta)/k$ is completely analogous). Then with probability $> \epsilon$ the number of T's between U_N/N^β and $[U_N + (c + \delta)/k]/N^\beta$ is at most $cN^\gamma + o(N^\gamma)$.

On the other hand, there exists A so large that with probability exceeding $1 - \epsilon/4$ we have $-A < U_N < U_N + (c + \delta)/k < A$ for all sufficiently large N. Given that this inequality holds, the probability is $> 1 - \epsilon/4$ that the number of T's between U_N/N^β and $[U_N + (c + \delta)/k]/N^\beta$ exceeds $[c + \delta/2]N^\gamma$ for all N sufficiently large. Hence the over-all probability that this number exceeds $(c + \delta/2)N^\gamma$ is greater than $(1 - \epsilon/4)^2 > 1 - \epsilon/2$, which gives the desired contradiction.

LEMMA 3. *Let W_N be the number of differences $Y_j - X_i$ satisfying $aN^{-\frac{1}{2}} < Y_j - X_i < bN^{-\frac{1}{2}}$, and let the distribution G of $Y_j - X_i$ have density g. Then*

$$(8) \qquad N^{-\frac{3}{2}}W_N = \rho(1 - \rho)g(0)(b - a) + R_N$$

where R_N tends to zero uniformly for $-A < a < b < A$.

PROOF. Let I_{ij} be equal to 1 if $aN^{-\frac{1}{2}} < Y_j - X_i < bN^{-\frac{1}{2}}$ and equal to zero otherwise. Then

$$E[N^{-\frac{3}{2}}W_N] = \rho(1 - \rho)(b - a)[G(bN^{-\frac{1}{2}}) - G(aN^{-\frac{1}{2}})]/(b - a)N^{-\frac{1}{2}}$$
$$\rightarrow \rho(1 - \rho)(b - a)g(0),$$

and the convergence is uniform in $-A < a < b < A$. To prove the desired result, it is enough to show that $\mathrm{Var}[N^{-\frac{3}{2}}W_N] = mnN^{-3}[\mathrm{Var}(I_{12}) + (n - 1)\mathrm{Cov}(I_{12}, I_{13}) + (m - 1)\mathrm{Cov}(I_{12}, I_{32})]$ tends to zero uniformly for a, b in any finite interval. Since $\mathrm{Var}(I_{12})$ is uniformly bounded, it only remains to show that the two covariances tend to zero uniformly. This follows, for example, from

$$|\mathrm{Cov}(I_{12}, I_{13})| \leqq P[|Y_2 - X_1| < AN^{-\frac{1}{2}} \text{ and } |Y_3 - X_1| < AN^{-\frac{1}{2}}]$$
$$+ P^2[|Y_1 - X_1| < AN^{-\frac{1}{2}}] \rightarrow 0$$

for the first covariance, and analogously for the other.

LEMMA 4. *As $N \rightarrow \infty$, the random variable $N^{\frac{1}{2}}(\Delta_L - \Delta)$ has a normal limiting distribution with mean $-K_{\alpha/2}/[12\rho(1 - \rho)]^{\frac{1}{2}}\int f^2(x) \, dx$ and variance $1/12\rho(1 - \rho)[\int f^2(x) \, dx]^2$.*

PROOF. Assuming without loss of generality that $\Delta = 0$, it is seen as in the proof of Lemma 1 that for any constant v, $P[N^{\frac{1}{2}}\Delta_L > v] = P[U(vN^{-\frac{1}{2}}) > mn - C_\alpha]$. The variables $N^{-\frac{3}{2}}[U(vN^{-\frac{1}{2}}) - \frac{1}{2}mn]$ have a normal limit distribution with mean

$$\lim_{N\to\infty} N^{-\frac{3}{2}}mn\{P[Y_1 - X_1 > vN^{-\frac{1}{2}}] - \frac{1}{2}\}$$

$$= \lim \rho(1 - \rho)N^{\frac{1}{2}}\int [F(y - vN^{-\frac{1}{2}}) - F(y)] \, dF(y) = -\rho(1 - \rho)v\int f^2(y) \, dy$$

and variance

$$\lim N^{-3}mn(m + n + 1)/12 = \rho(1 - \rho)/12.$$

Using Lemma 1, we therefore have

$$\lim P[N^{\frac{1}{2}}\Delta_L > v] = \lim P[U(vN^{-\frac{1}{2}}) - \tfrac{1}{2}mn > K_{\alpha/2}[mn(m + n)/12]^{\frac{1}{2}} + o(N^{\frac{3}{2}})$$

$$= 1 - \Phi\left(K_{\alpha/2} + v[12\rho(1 - \rho)]^{\frac{1}{2}} \int f^2(x)\, dx\right)$$

where Φ is the standard normal cumulative distribution function. This implies the desired result.

PROOF OF THEOREM 1. Take for the set $T_{1N}, \cdots, T_{NN}$ of Lemma 2 the differences $Y_j - X_i$, and for U_N and Z_N of Lemma 2 the statistics $N^{\frac{1}{2}}\Delta_L$ and $N^{\frac{1}{2}}\Delta_U$. Then

(i) the number of T's between $U_N/N^{\frac{1}{2}}$ and $Z_N/N^{\frac{1}{2}}$ is $2K_{\alpha/2}[\rho(1 - \rho)/12]^{\frac{1}{2}}N^{\frac{3}{2}} + o(N^{\frac{3}{2}})$, by Lemma 1;

(ii) the number of T's between $aN^{-\frac{1}{2}}$ and $bN^{-\frac{1}{2}}$ is $[\rho(1 - \rho)g(0)(b - a) + R_N] \cdot N^{\frac{3}{2}}$, with R_N tending to zero in probability uniformly for a, b in any finite interval by Lemma 3;

(iii) U_N is bounded in probability by Lemma 4.

Hence Lemma 2 is applicable with $\beta = \frac{1}{2}$, $\gamma = \frac{3}{2}$, $c = 2K_{\alpha/2}[\rho(1 - \rho)/12]^{\frac{1}{2}}$ and $k = \rho(1 - \rho)g(0)$. Since $g(0) = \int f^2(x)\, dx$, this establishes the desired result.

The length L' of the standard intervals (4) is given by $L' = 2K_{\alpha/2}S/[\rho(1 - \rho)N]^{\frac{1}{2}}$, and it therefore follows from Theorem 1 that the ratio of the squares of the lengths L'^2/L^2 tends in probability to the right hand side of (6). Alternatively, if the intervals (2) are based on N and the intervals (5) on N' observations, the ratio L'/L will tend in probability to one, provided (6) holds. Thus the right hand side of (6) is a reasonable measure of efficiency also when the comparison is made in terms of the lengths of the intervals.

It is interesting to compare the intervals (Δ_L, Δ_U) with the asymptotically distribution-free intervals, say $\Delta_L^* < \Delta < \Delta_U^*$, proposed in [1]. These were centered at the median $\text{med}(Y_j - X_i)$ of the mn differences $Y_j - X_i$ and had length $K_{\alpha/2}/[12NT_N\rho(1 - \rho)]^{\frac{1}{2}}$ where $(T_N)^{\frac{1}{2}}$ is any consistent estimator of $\int f^2(x)\, dx$. It is seen from the results of Section 5 of [1] and from Lemma 4 and Theorem 1 of the present paper that the joint limiting distribution of $(N^{\frac{1}{2}}(\Delta_L - \Delta), N^{\frac{1}{2}}(\Delta_U - \Delta))$ is the same as that of $(N^{\frac{1}{2}}(\Delta_L^* - \Delta), N^{\frac{1}{2}}(\Delta_U^* - \Delta))$, so that the two intervals have the same asymptotic behaviour. It is a consequence of this fact and the results of [1] that the right hand side of (6) also measures the efficiency of the present intervals in terms of the measure of accuracy

$$W(\Delta; \Delta_L, \Delta_U) = a(\Delta - \Delta_L)^2 + b(\Delta_U - \Delta)^2$$

in a sense made precise in Section 5 of [1].

We note finally that Theorem 1 proves

(9) $$N^{\frac{1}{2}}(\Delta_U - \Delta_L)[3\rho(1 - \rho)]^{\frac{1}{2}}/K_{\alpha/2}$$

to be a consistent estimator of $1/\int f^2(x)\, dx$. Thus (9) may be taken as the statistic $(T_N)^{\frac{1}{2}}$ required for (Δ_L^*, Δ_U^*).

The above results are easily extended to the one-sample problem. Let $Z_1, \cdots, Z_N$ be independent observations from a symmetric distribution F with median θ. Let $U'(\theta)$ be the number of averages $(Z_i + Z_j)/2$ with $i < j$ which exceed θ, and suppose that the two-sided symmetric level α test of the hypothesis $H'(\theta_0) : \theta = \theta_0$ accepts $H'(\theta_0)$ when

$$(10) \qquad C'_\alpha \leqq U'(\theta_0) \leqq \binom{N}{2} - C'_\alpha .$$

(This test is asymptotically equivalent to the Wilcoxon one-sample test, which could be used equally well throughout the argument.) If the ordered set of $\binom{N}{2}$ averages $(Z_i + Z_j)/2$ $(i < j)$ is denoted by $E^{(1)} < \cdots < E^{(M)}$ where $M = \binom{N}{2}$, then the confidence intervals obtained from (10) are seen to be

$$(11) \qquad \theta_L < \theta < \theta_U$$

where

$$(12) \qquad \theta_L = E^{(C'_\alpha)} \quad \text{and} \quad \theta_U = E^{(M+1-C'_\alpha)}.$$

For large N, one obtains in generalization of Lemma 1 the following approximation for C'_α.

LEMMA 1'. *The critical value* C'_α *is for large* N *equal to*

$$(13) \qquad C'_\alpha = N(N-1)/4 - K_{\alpha/2}(N-1)(N/12)^{\frac{1}{2}} + o(N^{\frac{3}{2}})$$

where $K_{\alpha/2}$ *has the same meaning as in Lemma* 1.

The proofs of Lemma 1', and of Lemmas 3' and 4' below will be omitted since they are exactly parallel to those of the corresponding unprimed lemmas. The efficiency results for the two-sample case also extend to the present problem, with the efficiency again being given by formula (6). The efficiency comparison in terms of length is based on the following.

THEOREM 1'. *If* $L' = \theta_U - \theta_L$ *is the length of the intervals defined by* (12) *and* (13), *then* $N^{\frac{1}{2}}L' \to K_{\alpha/2}/3^{\frac{1}{2}} \int f^2(x)\, dx$ *in probability as* $N \to \infty$.

The proof is completely analogous to that of Theorem 1, and is based on Lemma 2 and the following two lemmas.

LEMMA 3'. *Let* W'_N *be the number of pairs* $(Z_i + Z_j)/2$ *satisfying* $a/N^{\frac{1}{2}} < (Z_i + Z_j)/2 < b/N^{\frac{1}{2}}$ *and let the distribution* G *of* $(Z_i + Z_j)/2$ *have density* g. *Then*

$$(14) \qquad N^{-\frac{3}{2}}W'_N = g(0)(b-a)/2 + R_N$$

where R_N *tends to zero uniformly for* $-A < a < b < A$.

LEMMA 4′. *As $N \to \infty$, the random variable $N^{\frac{1}{2}}(\theta_L - \theta)$ has a normal limiting distribution with mean $-K_{\alpha/2}/(12)^{\frac{1}{2}} \int f^2(x) \, dx$ and variance $1/12 (\int f^2(x) \, dx)^2$.*

As a consequence of Theorem 1′, we note finally that for symmetric distributions F

$$(15) \qquad\qquad (3N)^{\frac{1}{2}}(\theta_U - \theta_L)/K_{\alpha/2}$$

is a consistent estimator of $1/\int f^2(x) \, dx$.

REFERENCE

[1] LEHMANN, E. L. (1963). Asymptotically nonparametric inference: An alternative approach to linear models. *Ann. Math. Statist.* **34** 1494–1506.