The Annals of Statistics
1973, Vol. 1, No. 6, 1200-1202

ON SEQUENTIAL CONFIDENCE INTERVALS BASED ON WILCOXON TYPE ESTIMATES[1]

BY M. S. SRIVASTAVA AND A. K. SEN

University of Toronto and
University of Illinois at Chicago Circle

For the location parameter family of distributions $F(x - \theta)$ under some regularity conditions, a confidence interval for θ of fixed width $2d$ and given confidence coefficient $1 - \alpha$ in the limit as d tends to zero is obtained using Hodges–Lehmann estimates based on Wilcoxon statistics. An upper bound on the average sample size is also given.

1. Introduction. Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population with cumulative distribution function (hereafter, cdf) $F(x - \theta)$. Under some regularity conditions on F, we wish to find a confidence interval I_N for θ such that (a) the length of $I_N \leqq 2d$ and (b) $\lim_{d\to 0} P\{\theta \in I_N\} \geqq 1 - \alpha$ where α and d are specified. Since no fixed-sample procedure can meet the above requirements, Geertsema [3] considered a sequential procedure in which N is a random variable and $N(d) \to \infty$ a.s. as $d \to 0$. He obtained confidence intervals based on sign and Wilcoxon tests (cf. Lehmann [5]) and showed them to be asymptotically efficient and consistent in the sense of Chow and Robbins [2]. The object of this note is to derive confidence intervals based on Hodges–Lehmann estimates using Wilcoxon statistics. We also give an upper bound for the average sample size $E(N)$.

2. Procedure based on Wilcoxon statistic. Let $\{X_n\}$ be a sequence of i.i.d. random variables with common cdf $F(x - \theta)$, where F is symmetric about 0 and has density f such that $\int f^2(x)\,dx < \infty$. Further let $Z_{n,1} \leqq Z_{n,2} \leqq \cdots \leqq Z_{n,p}$ be the $p \equiv \frac{1}{2}n(n+1)$ ordered averages $\frac{1}{2}(X_i + X_j)$, $i \leqq j$ and $i, j = 1, 2, \cdots, n$. Then the Hodges–Lehmann [4] estimate of θ is $\hat{Z}_n$ where $\hat{Z}_n$ is the median of $Z_{n,i}$'s, $i = 1, 2, \cdots, p$. We now define our stopping variable N as follows:

$$(1) \qquad N = \text{smallest integer } n \geqq n_0 \text{ such that}$$
$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} [I(-2d \leqq X_i - X_j \leqq 2d)] \geqq K_\alpha (n-1)(n/3)^{\frac{1}{2}} - n$$

where $I(A)$ denotes the indicator function of the Set A, $I(A) = 1$ if $X \in A$ and $I(A) = 0$ if $X \notin A$, and n_0 is so chosen as to make the right side of (1) positive and K_α is given by

$$\Phi(K_\alpha) = 1 - (\alpha/2)$$

where Φ is the standard normal cdf.

When sampling is stopped at $N = n$, choose

$$I_n = [\hat{Z}_n - d, \hat{Z}_n + d]$$

Received September 1971; revised February 1973.

[1] Research supported partially by the National Research Council of Canada.

as a confidence interval for θ. Clearly (a) is satisfied and (b) follows from (I) through (IV), below.

(I) Hodges and Lehmann [4, equation (9.2)] have shown that

$$P\{h(x-a) < \mu\} \leqq P\{\hat{Z}_n < a\} \leqq P\{h(x-a) \leqq \mu\}$$

with $h(x) =$ Number of pairs (i, j) with $1 \leqq i \leqq j \leqq n$ such that $X_i + X_j > 0$ and

$$\mu = \tfrac{1}{2}p \equiv \tfrac{1}{4}n(n+1)\,.$$

(II) $[n(n+1)]^{-1}h(x)$ is a U-statistic and can easily be shown to satisfy Anscombe's [1] condition (C2).

(III) Define a sequence $\{U_n\}$ by

$$U_n = \frac{2}{dn(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} [I(-2d \leqq X_i - X_j \leqq 2d)]\,.$$

Then $\{U_n\}$ forms a reverse martingale and hence as $n \to \infty$ and $d \to 0$

$$U_n \to 4 \int_{-\infty}^{\infty} f^2(x)\,dx \quad \text{a.s.}$$

(IV) Let $G(x)$ be the cdf of $\frac{1}{2}(X_i - X_j)$ $i \neq j$,

$$Y_n = \frac{n(n-1)[G(d) - \frac{1}{2}]}{[\sum_{i=1}^{n} \sum_{j=i+1}^{n} I(-2d \leqq X_i - X_j \leqq 2d)] + n}\,,$$

$$g(n) = n^{\frac{1}{2}} \qquad \text{and} \qquad t = \frac{K_\alpha}{[G(d) - \frac{1}{2}]} = \frac{K_\alpha}{d[(G(d) - \frac{1}{2})/d]}\,.$$

Then $Y_n > 0$ a.s. and $\lim_{n\to\infty} Y_n = 1$ a.s. from (III) above. Also $g(n) > 0$, $\lim_{n\to\infty} g(n) = \infty$, $\lim_{n\to\infty} [g(n)/g(n-1)] = 1$. Thus for each $t > 0$, N of (1) can be defined as

$$N = N(t) = \text{smallest} \quad n \geqq 1 \quad \text{such that} \quad Y_n \leqq g(n)/t\,.$$

Hence as in Lemma 1 of Chow–Robbins [2] it follows that N is well defined and non-decreasing as a function of t,

$$\lim_{t\to\infty} N = \infty \quad \text{a.s.} \qquad \text{and} \qquad \lim_{t\to\infty} E(N) = \infty$$

and

$$\lim_{t\to\infty} g(N)/t = 1 \quad \text{a.s.}$$

Next we give an upper bound for $E(N)$. By introducing a reverse stopping variable as in Simons [6], it can easily be shown that

$$E(N - n_0 + 1)^{-\frac{1}{2}} \geqq (K_\alpha^{\,2}/3)^{-1}(G(d) - \tfrac{1}{2})\,.$$

3. Remarks. REMARK 1. The stopping rule suggested in this paper is simpler than one suggested by Geertsema [3]. Geertsema suggested that sampling be stopped at the first integer $N \geqq n_0$ such that $Z_{n,a(n)} - Z_{n,b(n)} \leqq 2d$, where

$$a(n) \sim n(n+1)/4 + K_\alpha[n(n+1)(2n+1)/24]^{\frac{1}{2}}$$
$$b(n) \sim n(n+1)/4 - K_\alpha[n(n+1)(2n+1)/24]^{\frac{1}{2}}\,.$$

Thus, the computation requires the ranking of the averages $\frac{1}{2}(x_i + x_j)$, for every

n, whereas the present procedure requires only a count of those $x_i - x_j$ differences that lie between $-2d$ and $2d$. The latter is a considerably faster computation.

REMARK 2. The existence and the boundedness of the second derivatives of the cdf of $\frac{1}{2}(x_1 + x_2)$ in the neighborhood of θ is not required in our procedure in contrast to Geertsma's (1970).

REFERENCES

[1] ANSCOMBE, F. J. (1952). Large-sample theory of sequential estimation. *Proc. Cambridge Philos. Soc.* **48** 600–607.

[2] CHOW, Y. S. and ROBBINS, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.* **36** 457–462.

[3] GEERTSEMA, J. C. (1970). Sequential confidence intervals based on rank tests. *Ann. Math. Statist.* **41** 1016–1026.

[4] HODGES, J. L. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

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

[6] SIMONS, G. (1968). On the cost of not knowing the variance when making a fixed-width confidence interval for the mean. *Ann. Math. Statist.* **39** 1946–1952.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO, CANADA

CENTER FOR URBAN STUDIES
UNIV. OF ILLINOIS AT CHICAGO CIRCLE
CHICAGO, ILLINOIS 60660