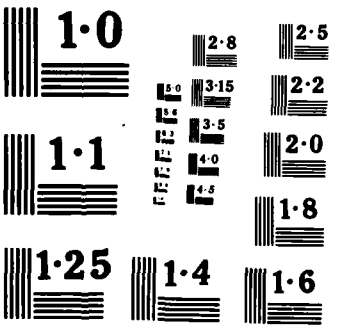
AD-A152 607 CONFIDENCE INTERVALS BASED ON INTERPOLATED ORDER STATISTICS(U) PENNSYLVANIA STATE UNIV UNIVERSITY PARK DEPT OF STATISTICS T P HETTMANSPERGER ET AL. MAR 85 1/1

UNCLASSIFIED TR-53 N00014-80-C-0741 F/G 12/1 NL


1·0
1·1
1·25
2·8
3·15
3·5
4·0
4·5
1·4
2·5
2·2
2·0
1·8
1·6

(2)

The Pennsylvania State University

Department of Statistics

University Park, Pennsylvania

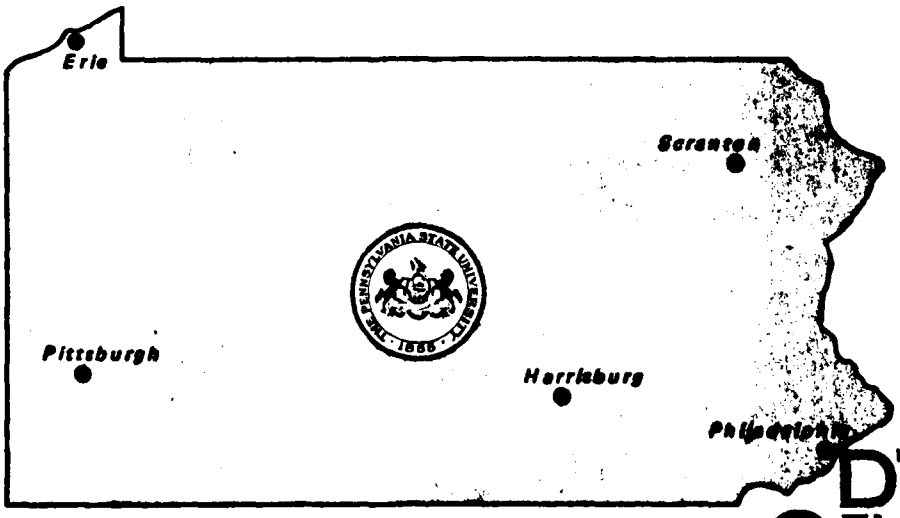
TECHNICAL REPORTS AND PREPRINTS

Number 53: March 1985

CONFIDENCE INTERVALS BASED ON INTERPOLATED ORDER STATISTICS

Thomas P. Hettmansperger*
The Pennsylvania State University and
La Trobe University

Simon J. Sheather
University of Melbourne



2

DEPARTMENT OF STATISTICS

The Pennsylvania State University
University Park, PA 16802 U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 53: March 1985

CONFIDENCE INTERVALS BASED ON INTERPOLATED ORDER STATISTICS

Thomas P. Hettmansperger*
The Pennsylvania State University and
La Trobe University

Simon J. Sheather
University of Melbourne

*Research partially supported by ONR Contract N00014-80-C0741

Abstract

Confidence intervals for the population median based on interpolating adjacent order statistics are presented. They are shown to depend only slightly on the underlying distribution. A simple, nonlinear interpolation formula is given which works well for a broad collection of underlying distributions.

Key words: Nonparametric, sign test.

1. Introduction

Suppose $X_1,\ldots,X_n$ is a random sample of size n from a distribution with absolutely continuous distribution function $F(x-\theta)$ and density $f(x-\theta)$. Further, suppose $F(0) = \frac{1}{2}$, uniquely, so that θ is the unique median; no shape assumption is imposed on F. Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics. Then the interval $[X_{(d)}, X_{(n-d+1)}]$ is a simple distribution-free confidence interval for θ. The confidence coefficient $\gamma = 1 - 2P(S<d)$ where S has a binomial distribution with parameters n and $\frac{1}{2}$. If S denotes the sign test statistic for testing $H_0:\theta = 0$ versus $H_A:\theta \neq 0$, then the interval corresponds to inverting the acceptance region of a size $\alpha = 2P(S<d)$ test. See Hettmansperger (1984a, Section 1.5).

This confidence interval is quite versitile since it makes no shape assumption on the underlying distribution, is easy to compute, and requires only a binomial table to establish the confidence coefficient. In Hettmansperger (1984b) we recommend using the interval to form the notches in a notched box plot and construct simple two sample tests based on comparing these intervals. The Minitab computing system uses these intervals in their box plot routine.

Because of the discreteness of the binomial distribution, for small to moderate sample sizes the available set of possible confidence coefficients is rather sparse. In this paper we consider the problem of interpolating adjacent order statistics to form confidence intervals with intermediate values of the confidence coefficients. The interpolated intervals are no longer distribution free in general; however, we will show that the confidence coefficient depends only slightly on the underlying F for a broad collection of distributions. Finally, we will show that linear

interpolation is not appropriate, and we will provide a simple interpolation formula that works well in most practical situations.

2. Properties of the Interpolated Intervals.

Suppose $[X_{(d)}, X_{(n-d+1)}]$ and $[X_{(d+1)}, X_{(n-d)}]$ are $\gamma_d = 1 - \alpha_d$ and $\gamma_{d+1} = 1 - \alpha_{d+1}$ confidence intervals for θ, respectively. Then, from the binomial distribution, we have

$$\frac{\alpha_{d+1}}{2} = \frac{\alpha_d}{2} + \binom{n}{d}\left(\frac{1}{2}\right)^n \tag{1}$$

or

$$\gamma_{d+1} = \gamma_d - 2\binom{n}{d}\left(\frac{1}{2}\right)^n .$$

This links the successive intervals based on d and d+1 .

Define, for $0 \le \lambda < 1$,

$$X_L = (1-\lambda)X_{(d)} + \lambda X_{(d+1)}$$

$$X_U = (1-\lambda)X_{(n-d+1)} + \lambda X_{(n-d)}$$

and let γ be the confidence coefficient for $[X_L, X_U]$. Then $\gamma_{d+1} \le \gamma < \gamma_d$ and we wish to establish the connection between λ and γ. Given γ, we will present a simple interpolation formula for finding λ; see (7) in Section 3. Note that

$$\begin{aligned}\gamma &= P(X_L \le \theta \le X_U) \\ &= 1 - P(\theta < X_L) - P(\theta > X_U) \\ &= 1 - \alpha_L - \alpha_U .\end{aligned}$$

Proposition 1. let $\sigma = \lambda/(1-\lambda)$. Then

$$\alpha_L = \frac{\alpha_{d+1}}{2} - (n-d)\binom{n}{d}\int_0^\infty [F(-\sigma y)]^d [1-F(y)]^{n-d-1} dF(y)$$
$$\alpha_U = \frac{\alpha_{d+1}}{2} - (n-d)\binom{n}{d}\int_{-\infty}^0 [1-F(-\sigma y)]^d [F(y)]^{n-d-1} dF(y) \quad . \tag{2}$$

Proof. Without loss of generality let $\theta=0$. Let D denote the set $\{(x,y): -\infty < x < y < \infty, (1-\lambda)x + \lambda y > 0\}$. Then, denoting the joint density of $X_{(d)}$ and $X_{(d+1)}$ by $f(x,y)$, we have

$$\alpha_L = P(0<X_L)$$

$$= \int_D\int f(x,y)\,dxdy$$

$$= \int_0^\infty\int_{-\sigma y}^{y} \frac{n!}{(d-1)!\,(n-d-1)!} [F(x)]^{d-1}[1-F(y)]^{n-d-1} f(x) f(y)\,dxdy$$

$$= \frac{n!}{(d-1)!\,(n-d-1)!}\,\frac{1}{d}\int_0^\infty [F(y)^d - F(-\sigma y)^d][1-F(y)]^{n-d-1} dF(y) \quad .$$

The formula for α_L now follows from the fact that

$$P(X_{(d+1)}>0) = \frac{n!}{d!\,(n-d-1)!}\int_0^\infty [F(y)]^d (1-F(y)]^{n-d-1} dF(y)$$

$$= \frac{\alpha_{d+1}}{2} \quad .$$

The formula for α_U follows in a similar way.

Proposition 2. Suppose f is symmetric about 0. Then

(i) $\alpha_L = \alpha_U$ (3)

(ii) if $\lambda = \frac{1}{2}$ it follows that

$$\alpha_L = \frac{\alpha_{d+1}}{2} - \frac{n-d}{n}\binom{n}{d}\left(\frac{1}{2}\right)^n$$
$$= \frac{\alpha_d}{2} + \frac{d}{n}\binom{n}{d}\left(\frac{1}{2}\right)^n \quad . \tag{4}$$

Proof. Part (i) follows at once from Proposition 1 since $F(\sigma y) = 1 - F(-\sigma y)$. In part (ii), note that $\lambda = \frac{1}{2}$ implies $\sigma = 1$, so that we have

$$\int_0^\infty [F(-y)]^d [1-F(y)]^{n-d-1} dF(y) = \int_0^\infty [1-F(y)]^{n-1} dF(y)$$

$$= \frac{1}{n}\left(\frac{1}{2}\right)^n .$$

The formulas in (ii) now follow from the result for α_L in Proposition 1 and the result in (1).

Given a desired γ, define the interpolation factor I by

$$I = \frac{\gamma_d - \gamma}{\gamma_d - \gamma_{d+1}} . \tag{5}$$

Note that I depends upon λ through γ so we will write $I(\lambda)$ when it is necessary to express this dependence.

Proposition 3. Let $\sigma = \lambda/(1-\lambda)$. Suppose f is symmetric about 0. Then

(i) $$I(\lambda) = 1 - (n-d)2^n \int_0^\infty [F(-\sigma y)]^d [1-F(y)]^{n-d-1} dF(y) \tag{6}$$

$I(0) = 0$, $I(\frac{1}{2}) = d/n$ and $I(\lambda) \to 1$ as $\lambda \to 1$.

(ii) If F is sufficiently regular so that differentiation can be caried out under the integral and if $f'(x) > 0$ for $x \leq 0$, then $I(\lambda)$ is a continuous and strictly increasing convex function of λ.

Proof. Using $\gamma = 1-2\alpha$, from (5) and (1), we have $I = [\alpha_{d+1}/2 - \alpha_d/2]/\binom{n}{d}\left(\frac{1}{2}\right)^n$. The formula for I now follows from Propositions 1 and 2 by substitution. The limit follows by the dominated convergence theorem since $F(-\sigma y) \to 0$ as $\sigma \to \infty$ for $y > 0$. Part (ii) follows by verifying that $I'(\lambda) > 0$ and $I''(\lambda) > 0$.

This proposition shows at once that linear interpolation is inappropriate. If we used linear interpolation then $I(\frac{1}{2})$ must be equal to $\frac{1}{2}$. However, $I(\frac{1}{2}) = d/n$ which is less than $\frac{1}{2}$. For example with $n=10$, $d=2$, $\gamma_d = .9786$, $\gamma_{d+1} = .8907$, we have $I = d/n = .2$ and $\gamma = .9610$, corresponding to $\lambda = \frac{1}{2}$. Linear interpolation yields .9347. As a crude approximation take $d \doteq n/2 - Zn^{-\frac{1}{2}}/2 + .5$, where Z is the $\alpha_d/2$ quantile from the standard normal distribution, then

$$\frac{d}{n} \doteq \frac{1}{2} - \frac{Z}{2n^{\frac{1}{2}}} + \frac{1}{2n} < \frac{1}{2} .$$

Further, since $I(\frac{1}{2}) = d/n$, we have an additional distribution-free interval when f is symmetric. This helps in the search for an interpolation formula since the curve for $I(\lambda)$ must pass through the ordinates 0, d/n, and 1, at least for a symmetric f. In principle, given $\lambda \neq \frac{1}{2}$ we would need to specify F to find I according to (6). In practice, we wish to specify I, through γ, and find λ. From Part (ii) there exists a strictly increasing concave curve that relates λ to I but is generally impossible to find because of the complexity of (6).

In the next section we find $I(\lambda)$ for several different distributions. We show that the curves are quite close to one another.

We then select one which yields a particularly simple formula for $I(\lambda)$, (it results when F is the double exponential distribution), and invert it to provide a formula for λ in terms of I.

3. Examples and a Recommendation

In this section we provide explicit formulas for $I(\lambda)$ for underlying uniform and double exponential distributions and an asymmetric distribution formed by piecing together double exponential and logistic distributions. We also provide numerical results, based on numerical integration, for the normal and Cauchy distributions. The numerical integrations were carried out using the method of Donker and Piessens (1975).

Numerical examples are given in Tables 1 and 2.

Example 1. The double exponential distribution. The distribution function is given by $F(x) = 2^{-1}\exp(x)$ if $x < 0$, and $1 - 2^{-1}\exp(-x)$ if $x \geq 0$. In (6) replace $F(-\sigma y)$ by $1 - F(\sigma y)$ and then

$$\int_0^\infty [1-F(\sigma y)]^d [1-F(y)]^{n-d-1} dF(y) = \left(\frac{1}{2}\right)^n \int_0^\infty \exp\{-y(d\sigma+n-d)\}dy$$

$$= [n+d(\sigma-1)]^{-1}\left(\frac{1}{2}\right)^n$$

Hence, from (6), $I(\lambda) = d\sigma/[n+d(\sigma-1)]$. Recalling that $\sigma = \lambda/(1-\lambda)$ we find

$$\lambda = \frac{(n-d)I}{d + (n-2d)I} . \qquad (7)$$

As a final remark, note that if we take a curve fitting approach and try to find a concave curve through $(0,0)$, $(d/n,\frac{1}{2})$ and $(1,1)$, then (7) results when we fit $\lambda = aI/(b+cI)$. Polynomial fits were not very satisfactory and (7) represents a simple ratio of linear polynomials.

Example 2. The uniform distribution on $(-1,1)$. The distribution function is given by $F(x) = 0$ if $x < -1$, $2^{-1}(x+1)$ if $|x| \leq 1$, and 1 if $x > 1$. Note also that $F(-\sigma y) = 0$ if $y > \sigma^{-1}$, $2^{-1}(-\sigma y+1)$ if $-\sigma^{-1} \leq y \leq \sigma^{-1}$, and 1 if $y < -\sigma^{-1}$. Now, using a binomial expansion on $(1-\sigma y)^d$ and the beta integral, we have, when $\sigma^{-1} \geq 1$ $(\lambda \leq \frac{1}{2})$,

$$\int_0^\infty [F(-\sigma y)]^d [1-F(y)]^{n-d-1} dF(y) = \left(\frac{1}{2}\right)^n \int_0^1 (1-\sigma y)^d (1-y)^{n-d-1} dy$$

$$= \left(\frac{1}{2}\right)^n \frac{d!\,(n-d-1)!}{n!} \sum_{j=0}^{d} (-\sigma)^j \binom{n}{d-j} \quad (8)$$

When $\sigma^{-1} \leq 1$ $(\lambda \geq \frac{1}{2})$ we have

$$\int_0^\infty [F(-\sigma y)]^d [1-F(y)]^{n-d-1} dF(y) = \left(\frac{1}{2}\right)^n \int_0^{\sigma^{-1}} (1-\sigma y)^d (1-y)^{n-d-1} dy$$

$$= \sigma^{-1}\left(\frac{1}{2}\right)^n \int_0^1 (1-t)^d (1-\sigma^{-1}t)^{n-d-1} dt$$

$$= \sigma^{-1}\left(\frac{1}{2}\right)^n \frac{d!\,(n-d-1)!}{n!} \sum_{j=0}^{n-d-1} (-\sigma^{-1})^j \binom{n}{n-d-j-1}$$

$$\vdots \quad (9)$$

The formulas (8) and (9) can then be used to calculate $I(\lambda)$ for various values of λ.

Example 3. Pieced together double exponential and logistic distributions. The distribution function is given by $F(x) = 2^{-1}\exp(x/\tau)$ if $x < 0$, and $[1 + \exp(-x)]^{-1}$ if $x \geq 0$. If $\tau = 2$, the density function is continuous, the first quartile is -1.39, and the third quartile is 1.1. If $\tau = 10$, there is a jump in the density at 0, the first quartile is -6.9, and the third quartile is 1.1. Because of the asymmetry we must evaluate α_L and α_U separately. See formulas (2).

We have

$$\int_0^\infty [F(-\sigma y)]^d [1 - F(y)]^{n-d-1} dF(y)$$

$$= \left(\frac{1}{2}\right)^d \int_0^\infty \exp\left[-y\left(\frac{\sigma d}{\tau} + n - d\right)\right]\{1 + \exp(-y)\}^{-n+d-1} dy \qquad (10)$$

and

$$\int_{-\infty}^0 [1 - F(-\sigma y)]^d [F(y)]^{n-d-1} dF(y)$$

$$= \left(\frac{1}{2}\right)^{n-d} \frac{1}{\sigma\tau} \int_0^\infty \exp\left[-y\left(d + \frac{n-d}{\sigma\tau}\right)\right]\{1 + \exp(-y)\}^{-d} dy \; . \qquad (11)$$

By making the change of variable $u = \exp(-y)$ both integrals can be reduced to the form

$$\int_0^1 u^{N-1}(1+u)^{-M} du = \left(\frac{1}{2}\right)^N \sum_{j=0}^{\infty} \binom{M-N-1}{j} \left(-\frac{1}{2}\right)^j (N+j)^{-1} \; . \qquad (12)$$

See Gradshten and Ryzhik (1965 p.285). Using (12), (10) can be written as

$$\left(\frac{1}{2}\right)^{\frac{\sigma d}{\tau}+n-d} \sum_{j=0}^{\infty} (j!2^j)^{-1} \left[\frac{\sigma d}{\tau}\left(\frac{\sigma d}{\tau}+1\right)\cdots\left(\frac{\sigma d}{\tau}+j-1\right)\right]\left(\frac{\sigma d}{\tau}+n-d+j\right)^{-1}$$

and (11) can be written as

$$\left(\frac{1}{2}\right)^{d+\frac{n-d}{\sigma\tau}} \sum_{j=0}^{\infty} (j!2^j)^{-1} \left|\left(\frac{n-d}{\sigma\tau}+1\right)\cdots\left(\frac{n-d}{\sigma\tau}+j\right)\right| \left(d+\frac{n-d}{\sigma\tau}+j\right)^{-1} .$$

These infinite series are straightforward to approximate. For large j the (j+1)st term is roughly one-half of the jth term. This means that the tail of the series is roughly equal to the last term retained. The examples are calculated accurately to four places.

For the normal, logistic and Cauchy distributions we used numerical integration as mentioned previously. As an illustration we take n=10, d=2, γ_d=.9786, γ_{d+1}=.8907. In Table 1 we show $I(\lambda)$ for λ = .1 (.1) .9 and $\gamma = \gamma_d - (\gamma_d - \gamma_{d+1})I$ in parentheses. Linear interpolation results are provided for comparison.

- Table 1 about here -

Note in Table 1 how close γ is for all λ and for the spread of distributions uniform to Cauchy. The logistic distribution was indistinguishable from the normal distribution so it was not included in the table.

If the underlying distribution can be supposed to be symmetric then we recommend using formula (7) to determine λ from I. For the example considered here, if we want a 95% confidence interval then from (5) I = .3254 and from (7) λ = .66. Hence $X_L = .34X_{(2)} + .66X_{(3)}$ and $X_U = .34X_{(9)} + .66X_{(8)}$ provide the 95% confidence interval.

In Table 2 we illustrate the n=10, d=2 case for the asymmetric distribution in Example 3. The table shows the lower and upper tails, α_L and α_U, and then compares $\gamma = 1 - \alpha_L - \alpha_U$ to the γ calculated from the double exponential example.

- Table 2 about here -

Table 2 shows that mild asymmetry does not matter much and we would still use (7). In the pathological case, τ=10, the results were surprisingly close even though the two tails differed by quite a bit. Further, for this extreme case, linear interpolation is at least twice as far from γ as (7).

Hence, we conclude that for most practical situations (7) provides an accurate interpolation formula.

4. Acknowledgement

The research of the first author was partially supported by ONR Contract N00014-80-C0741. We would like to thank Ian Robinson for providing the numerical integration subroutine.

References

De Donker, E. and Piessens, R. (1975). Automatic Computation of Integrals with Singular Integrand over a Finite or Infinite Interval. Report TW22, Applied Maths. and Progr. Divn, Kath. Univ. Leuven.

Gradshten, I.S. and Ryzhik, I.M. (1965). Tables of Integrals, Series and Products. New York : Academic Press.

Hettmansperger, T.P. (1984a). Statistical Inference Based on Ranks. New York : John Wiley.

Hettmansperger, T.P. (1984b). Two-Sample Inference Based on One-Sample Sign Statistics. Appl. Statist., 33, 45-51.

Table 1. Interpolation Factors and Confidence Coefficients.

λ	DE*	U	N	C	LINEAR
.1	.027(.976)**	.023(.977)	.025(.976)	.026(.976)	.1(.970)
.2	.059(.973)	.052(.974)	.055(.974)	.057(.974)	.2(.961)
.3	.096(.970)	.091(.971)	.092(.971)	.094(.970)	.3(.952)
.4	.143(.966)	.137(.967)	.139(.966)	.141(.966)	.4(.943
.5	.200(.961)	.200(.961)	.200(.961)	.200(.961)	.5(.935)
.6	.273(.955)	.282(.954)	.280(.954)	.275(.954)	.6(.926)
.7	.369(.946)	.396(.944)	.388(.944)	.373(.946)	.7(.917)
.8	.501(.935)	.553(.930)	.536(.931)	.506(.934)	.8(.908)
.9	.692(.918)	.753(.912)	.736(.914)	.694(.918)	.9(.899)

* DE = Double exponential, U = Uniform, N = Normal, C = Cauchy.

** The number in parentheses is the confidence coefficient.

Table 2. Confidence Coefficients in Asymmetric Case

$\tau = 2$

λ	α_U	α_L	γ	DE
.1	.0119	.0117	.976	.976
.2	.0133	.0130	.974	.973
.3	.0151	.0146	.970	.970
.4	.0173	.0165	.966	.966
.5	.0201	.0189	.961	.961
.6	.0237	.0220	.954	.955
.7	.0284	.0262	.945	.946
.8	.0346	.0320	.933	.935
.9	.0430	.0407	.916	.918

$\tau = 10$

λ	α_U	α_L	γ	DE
.1	.0163	.0109	.973	.976
.2	.0220	.0112	.967	.973
.3	.0275	.0115	.961	.970
.4	.0325	.0120	.956	.966
.5	.0372	.0126	.950	.961
.6	.0414	.0135	.945	.955
.7	.0452	.0149	.940	.946
.8	.0487	.0175	.934	.935
.9	.0518	.0236	.925	.918

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 53	2. GOVT ACCESSION NO.: AD-A152 667
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): CONFIDENCE INTERVALS BASED ON INTERPOLATED ORDER STATISTICS	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Thomas P. Hettmansperger, Penn State & La Trobe Universities Simon J. Sheather, University of Melbourne	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0741
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics The Pennsylvania State University University Park, Pa 16802.	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR042-446
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistical and Probability Program Code 436 Arlington, Va. 22217.	12. REPORT DATE: March 1985 13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nonparametric, Sign Test

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Confidence intervals for the population median based on interpolating adjacent order statistics are presented. They are shown to depend only slightly on the underlying distribution. A simple, nonlinear interpolation formula is given which works well for a broad collection of underlying distributions.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

5-85

DTIC