NBER WORKING PAPER SERIES

A COMPARISON OF TWO SIMPLE METHODS FOR OBTAINING ROBUST CONFIDENCE INTERVALS FOR A LOCATION PARAMETER

Richard W. Hill*

Working Paper No. 84

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

May 1975

Preliminary: not for quotation

* NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

Abstract

In this paper we study two methods for finding confidence limits for the simple median. One method is the new parametric procedure based on the sign test, and the other is derived in the paper. The two methods are compared asymptotically and also for small samples.

Contents

I. Introduction

Recently there has been a great deal of interest in robust estimators of location and regression parameters. Many of the better estimators are not really suitable for hand calculation (see Andrews et al. (1972)); the sample median, however, is a reasonable estimator of location that is readily computed by hand. Once a point estimate has been computed, it is natural to seek confidence limits for the estimate; the purpose of this paper is to examine two alternative methods for setting robust confidence intervals for the sample median.

II. The Procedures

The first procedure we will examine is the nonparametric procedure based on the sign test. This is a well known procedure, described in many places (e.g. Thompson (1936), Scheffe (1943), Noether (1949), Dixon (1953), Fraser (1957)). Let $X_1,\ldots,X_n$ denote the order statistics of the sample and $\tilde{X}$ the sample median. The nonparametric confidence interval of level α for $\tilde{X}$ is (X_r, X_{n-r+1}), where r is chosen so that the probability of fewer than r successes in a binomial with parameters $(\frac{1}{2},n)$ is less than or equal to $\frac{\alpha}{2}$. Nair (1940) and McKinnon (1964) have tabled r as a function of n for the 5% and 1% levels. It turns out that for the 5% level r=1 until n=9, and for the 1% level r=1 until n=12. Thus the nonparametric procedure cannot be considered robust for n less than 9 or 12 (since gross errors may affect the length of the interval violently for small n). Incidentally, it is worth noting that the formulas based on the normal approximation to the binomial, $r = \frac{n}{2} - \sqrt{n}$ for the 5% interval, and

$r = \frac{n}{2} - 1.3\sqrt{n}$ for the 1% interval (where r is rounded to the closest integer) (see Nair-1940) work well for $n \geq 10$ (they are conservative) and can safely be used if tables are not available.

The second procedure we shall examine is robust for $n \geq 5$. Let $r' = \frac{n+2}{4}$. Put

$$r = \begin{cases} [r'] & \text{if} \quad r' - [r'] \leq .5 \\ [r'] + 1 & \text{otherwise} \end{cases} .$$

Then let

$$MS = \begin{cases} X_r - X_{m-r+1} & \text{if} \quad r' - [r'] \neq .5 \\ \frac{X_r + X_{r+1} - X_{m-r+1} - X_{m-r}}{2} & \text{otherwise.} \end{cases}$$

The random quantity MS is sometimes called the interquartile range, or the midspread (Tukey-1970). We define our confidence interval by

$$(\tilde{X} - t'(\alpha,n)\frac{MS}{\sqrt{n}} , \tilde{X} + t'(\alpha, n)\frac{MS}{\sqrt{n}}) , \tag{1}$$

where $t'(\alpha,n)$ is chosen to achieve the desired level. In other words, we mimic student's interval, but with robust estimates of scale and location. This interval has been studied in the past: Birnbaum (1970) derived distribution free bounds for $t'(\alpha,n)$, which are unfortunately far too conservative; David and Johnson (1956) studied this interval when sampling from a normal distribution; and Weisberg (1973) studied generalizations of this interval. There are two main problems with this interval: 1) clearly $t'(\alpha,n)$ will vary from distribution to distribution. In other words a single set of $t'(\alpha,n)$ values will not insure α level intervals for all distributions. To some extent Student's interval has the same problem, although is widely used. 2) For any specific population, how do we determine $t'(\alpha,n)$? The exact determination of $t'(\alpha,n)$ involves a 3 dimensional integral, which can pose severe numerical difficulties. Both problems can be side-stepped simultaneously by choosing $t'(\alpha,n)$ so that level α is approximately achieved at the Gaussian distribution. Since the median is relatively more efficient in non-Gaussian situations, we would hope that this method will insure a conservative interval. As a first step toward approximating $t'(\alpha,n)$, we ignore n and use asymptotic theory.

III. Asymptotic Theory

Let $N(\mu,\sigma^2)$ denote the normal law with mean μ and variance σ^2. Let $F(\cdot)$ denote the cdf of the symmetric distribution from which we are sampling, and let $F'(\cdot) = f(\cdot)$. Then it is easily shown (Cramer-1946) that

$$\sqrt{n}\ \tilde{X} \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{[2\ f(0)]^2}\right) \tag{2}$$

By symmetry and the law of large numbers, we also have

$$MS \xrightarrow{P} 2F^{-1}(.75)$$

It follows at once that

$$\frac{\sqrt{n}\ \tilde{X}}{MS} \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{[4\ f(0)F^{-1}(.75)]^2}\right) \tag{3}$$

If Z_α denotes the $\frac{\alpha}{2}$ percentage point of the normal distribution, this suggests that we can calibrate our interval for the normal distribution by taking

$$t'(\alpha,n) = Z_\alpha \frac{1}{4\ \phi(0)\Phi^{-1}(.75)} \overset{\sim}{=} Z_\alpha\ .94 \tag{4}$$

The factor .94 is sufficiently close to 1 that we can discard it, so we suggest the approximation

$$t'(\alpha,n) = t(\alpha,n), \tag{5}$$

where $t(\alpha,n)$ denotes the $\frac{\alpha}{2}$ percentage point of Student's t distribution on n degrees of freedom.

In order to compare the two intervals, we shall consider $\sqrt{n}$ (expected length). for the interval (1) this is

$$- 4\ t'(\alpha,n)\ E\, X_{r_2}\ , \tag{6}$$

where $r_2 = \frac{n+2}{4}$ rounded to the nearest integer. (To avoid trivial complications, we will avoid the case when $\frac{n+2}{4} - [\frac{n+2}{4}] = .5$).

For the nonparametric interval, $\sqrt{n}$ (expected length) is

$$- 2 \sqrt{n}\ E\, X_{r_1}\ , \tag{7}$$

where r_1 is tabled, and approximately $r_1 = \frac{n}{2} - \frac{Z_\alpha}{2} \sqrt{n}$.

Clearly the asymptotic length of interval (1) is

$$4\ Z_\alpha\ F^{-1}(.75), \tag{8}$$

and it's asymptotic level is

$$1- 2\ \Phi(Z_\alpha\ 4\ F^{-1}(.75)\ f(0)\ ) \tag{9}$$

It is instructive to compute the asymptotic length of the nonparametric interval. We want to determine

$$\lim_{n\to\infty} - 2\sqrt{n}\ F^{-1}\left(\frac{\frac{n}{2} - \frac{Z_\alpha}{2}\sqrt{n}}{n}\right) = \lim_{n\to\infty}\ 2\sqrt{n}\ F^{-1}\left(\frac{1}{2} + \frac{Z_\alpha}{2}\frac{1}{\sqrt{n}}\right) .$$

Expanding F^{-1} in a Taylor series about $\frac{1}{2}$ and using the relation

$\frac{d}{dp}\ F^{-1}(p) = \frac{1}{f(F^{-1}(p))}$, we find, assuming smoothness and symmetry,

$$F^{-1}(\tfrac{1}{2}+h) = h\,\frac{1}{f(0)} + \frac{h^3}{6}\left(-\,\frac{f''(0)}{[f(0)]^4}\right) + \frac{h^5}{120}\left(10\,\frac{[f''(0)]^2}{[f(0)]^7} - \frac{f''''(0)}{[f(0)]^6}\right) + \dots \qquad (10)$$

Hence

$$2\sqrt{n}\ F^{-1}\left(\frac{1}{2} + \frac{Z_\alpha}{2}\,\frac{1}{\sqrt{n}}\right) = 2\sqrt{n}\ \frac{Z_\alpha}{2}\,\frac{1}{\sqrt{n}}\,\frac{1}{f(0)} + 0\left(\frac{1}{n}\right)$$

and

$$\lim_{n\to\infty}\ 2\sqrt{n}\ F^{-1}\left(\frac{1}{2} + \frac{Z_\alpha}{2}\ \frac{1}{\sqrt{n}}\right) = \frac{Z_\alpha}{f(0)} \qquad (11)$$

Variations of ezpression (11) are well known, and by comparing (11) with (2), we see that the nonparametric interval is asymptotically the shortest possible α level interval for the median, since asymptotically the optimal α level interval must have length $2\ Z_\alpha\,\frac{1}{2f(0)} = \frac{Z_\alpha}{f(0)}$. The asymptotic efficiency of interval (1) relative to the nonparametric interval is given by

$$\frac{4\ Z_\alpha\ F^{-1}\ (.75)}{\dfrac{Z_\alpha}{f(0)}} = 4\ f(0)\ \ F^{-1}\ (.75) \qquad (12)$$

(from (8) and (11)).

Expressions (12) and (9) show that the interval (1) will in fact be conservative (at least asymptotically) so long as it is asymptotically inefficient with respect to the nonparametric interval. To examine this further, we approximate

We evaluate (12) for various distributions:

% Asymptotic Inefficiency of Interval (1) to Nonparametric

Normal	Cauchy	Double Exponential
8	27	40

Next we consider the contaminated normal distribution, where we have

$$F(\mathbf{x}) = (1-p)\,\Phi\,(x) + p\,\Phi\,(\tfrac{x}{K}) \quad (0 \le p \le \tfrac{1}{2}) \quad .$$

The equation

$(1-p)\,\Phi\,(x) + p\,\Phi\,(\frac{x}{k}) = .75$ is easily solved by Newton's method and leads to the inefficiencies given in Table I.

These results suggest that in moderate contamination with fat tails, the interval (1) is probably conservative and not very inefficient asymptotically, although for a large contamination fraction it may become very inefficient. ($p = .25$, $k = .01$ is equivalent to $p = .75$, $k = 100$, from the definition of the contaminated normal).

IV. Some Finite Sample Results

Gross (1971) performed a Monte Carlo experiment for a wide variety of estimators. Among other quantities, he estimated $t'(\alpha,n)$ for $\alpha = .05$ under various sampling distributions. His results are given in Table II. David and Johnson (1956) considered the statistic $\frac{\tilde{X}}{MS}$, and found approximations to the percentage points of its distribution. From their results, we can derive the values of $t'(\alpha,n)$ for $\alpha = .05$ and $\alpha = .02$ under normality. These results are given in Table III, together with the ratio of t' to Student's t.

Table I

% Inefficiency for Contaminated Normal

K=	.01	.1	3	5	10	100
p=.01	112	16	8	8	8	8
.1	926	80	9	9	9	10
.25	1671	123	10	12	14	15
.5	90	35	15	22	35	91

Table II

$t'(\alpha,n)$ for $\alpha = .05$

Distribution	G	S/4	S	C	10G/20	10G/4	3S/4
n = 10	2.19	2.10	1.91	1.76	-	-	-
20	1.96	-	-	1.65	-	-	-
40	1.88	1.82	-	1.54	1.88	1.73	1.78
∞	1.82	-	-	1.54	-	-	-

Asymptotic values are from (9). G stands for normal, C Cauchy, S normal divided by independent uniform, S/4 a mixture of 25% S and 75% normal, 3S/4 same as S/4 except that each observation from S is multiplied by 3, 10G/20 exactly 1 observation in 20 from G is multiplied by 10, 10G/4 exactly 5 observations in 20 from G are multiplied by 10.

Table III

$t'(\alpha,n)$ for the Gaussian

	5%		2%	
	$t'(\alpha,n)$	t/t'	t'	t/t'
n=11	2.066	.940	2.905	1.069
15	1.991	.934	2.618	1.006
19	1.953	.933	2.498	.984
23	1.928	.932	2.427	.971
27	1.907	.930	2.380	.962
31	1.899	-	2.350	-
35	1.887	-	2.325	-
39	1.880	-	2.304	-
43	1.875	-	2.284	-
47	1.872	-	2.276	-
51	1.864	-	2.264	-
∞	1.821	.930	2.161	.930

Table IV

$t'(\alpha,n)$ for the Gaussian

	5%			1%		
	Crude	Swindle	t	Crude	Swindle	t
n=5	4.76	4.64	2.57	11.09	11.14	4.03
7	1.98	1.96	2.37	3.25	3.22	3.50
9	2.70	2.80	2.26	5.01	4.84	3.25
11	1.92	1.92	2.20	2.90	2.85	3.11
21	2.17	2.12	2.08	3.23	3.01	2.83
41	1.92	1.96	2.02	2.81	2.70	2.70

"Crude"refers to estimates based on the observed percentage points.

The results presented in Table II and III suggest that using $t'(\alpha,n) = t(\alpha,n)$ is indeed a conservative procedure, for $\alpha = .05$ and $n \geq 10$, but Table II suggests that this need not hold for $n < 10$ or $\alpha = .01$. To examine this, I conducted a Monte Carlo experiment under the normal distribution for $n = 5, 7, 9, 11, 21, 41$ (I used respectively 6000, 4000, 3000, 2000, 1000 and 700 replications). Since the results are merely meant to provide a crude approximation to $t'(\alpha,n)$, I will not describe the experiment in detail (It was carried out in the TROLL system using a Box-Mueller normal generator. A standard variance reduction technique (sometimes referred to as "swindling") was employed (Holland 1975). The results are presented in Table IV, together with the relevant t values. The irregular behavior as n increases is not due to sampling error, but to the irregularities in the definition of MS. From Table IV we see that the approximation $t'(.05,n) = t(.05,n)$ breaks down for $n < 10$, and that the approximation is never very good for $\alpha = .01$. We propose the following rules for $t'(\alpha,n)$:

For $\alpha = .05$: if $n \geq 10$ use Student's t. If $n < 10$, use $7.5 - \frac{n}{2}$.

for $\alpha = .01$: first, find $t'(.05,n)$. Then set

$$t'(.01,n) = \begin{matrix} 2\ t'(.05,n) & \text{if } n < 15 \\ 1.5\ t'(.05,n) & \text{if } n \geq 15 \end{matrix} \quad .$$

Now that we have approximate values of $t'(\alpha,n)$ even for small n, we examine the lengths of the two intervals for small n. As already noted, the 5% non-parametric interval is robust only for $n \geq 10$, and the 1% only for $n \geq 12$. Our first results for the 5% intervals are again taken from Gross (1971). They are presented in Table V.

Table V

Relative inefficiency of expected lengths of interval (1) to the nonparametric interval $[100 (\frac{\text{expected length of (1)}}{\text{expected length of NP}} - 1)]$.

	G	S/4	S	C	10G/20	10G/4	3S/4
n=10	20	0	-18	-20	-	-	-
20	11	-	-	17	14	17	15
40	9	14	-	26	-	-	-
∞	8	-	-	27	-	-	-

Asymptotic results are from (12). The symbols have the same meaning as in Table II.

These results suggest that for n=10, the interval (1) may be superior in fat-tailed situations. To examine this further, we turn to the contaminated normal distribution. Gastwirth and Cohen (1970) have tabled expected values of order statistics for some contaminated normal distributions. The results of Table VI are dervied from their table.

Table VI

% inefficiency for contaminated normal, K=3

	p=.01		p=.1	
	5%	1%	5%	1%
n=10	- 9	13	-11	- 1
14	-16	25	-17	5
18	- 1	13	- 5	24
20	3	-	3	-
∞	8	23	9	24

Average length of the nonparametric interval

	p=.01		p=.1	
	5%	1%	5%	1%
n=10	6.41	10.09	7.21	13.00
14	6.82	9.15	7.50	10.27
18	5.64	7.30	6.39	7.90
20	5.32	8.32	5.74	9.11

Average length of the interval (1)

	p=.01		p=.1	
	5%	1%	5%	1%
n=10	5.88	11.78	6.41	12.81
14	5.75	11.53	6.14	12.39
18	5.62	8.42	6.05	9.08
20	5.62	-	6.05	-

Again, we see that for small n and contaminated situations, the interval (1) may be shorter than the nonparametric interval.

V. Example

As an example, we compute the two intervals for a specific set of data, together with Student's interval. The data are differences in the weights of matched pairs of rats undergoing a certain experiment, and were collected at Harvard. The order statistic is -75, -54, -51, 0, 5, 12, 14, 15, 16, 17, 22, 22, 29, 38, 41 with n = 15.

We find

$\tilde{X} = 15$, $MS = 22-0 = 22$, $\sqrt{n} \cong 3.87$

The 5% level t value is 2.13 so the interval (1) becomes 15 ± 12 = (3, 27) The nonparametric interval is (0,22).

The sample mean is 3.4 and the sample standard deviation is 33.7 so Student's interval is 3.4 ± 9 = (-5.6, 9.4). A normal plot shows that the data are clearly not a sample from a normal distribution, so in this case, either robust interval is preferable to Student's interval.

VI. Conclusion

It is not necessary to go to a great deal of trouble to get a robust confidence interval with reasonable properties. For $n \geq 20$, the nonparametric interval based on $r = \frac{n}{2} - \sqrt{n}$, or $r = \frac{n}{2} - 1.3\sqrt{n}$ is very good. For $10 < n < 20$ either the nonparemtric interval, or $\tilde{X} \pm t'(\alpha,n)MS$ is reasonable. If $5 \leq n \leq 10$, the interval $\tilde{X} \pm t'(\alpha,n)MS$ is plausible. A simple approximation to $t'(\alpha,n)$ is:

$$t'(.05,n) = \begin{cases} \text{Student's } t \text{ for } n \geq 10 \\ 7.5 - \frac{n}{2} \qquad \text{otherwise} \end{cases}$$

and

$$t'(.01,n) = \begin{cases} 2 \quad t'(.05,n) \quad \text{if } n < 15 \\ 1.5\ t'(.05,n) \quad \text{otherwise.} \end{cases}$$

References

Andrews, et al. (1972), *Robust Estimates of Location*, Princeton University Press.

Birnbaum, F.W. (1970), On a Statistic Similar to Student's t. In *Nonparametric Techniques in Statistical Inference*, ed. Madan Lal Puri, Cambridge University Press.

Cramer, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

David, F.N. and Johnson, N.L. (1956), Some Tests of Significance with Ordered Variables. *Journal of the Royal Statistical Society*, B, Vol. 18, p. 1.

Dixon, W.J. (1953), Power Functions of the Sign Test and Power Efficiency for Normal Alternatives. *Annals of Mathematical Statistics*, Vol. 22, p. 467.

Fraser, D.A.S. (1957), *Nonparametric Methods in Statistics*, John Wiley and sons.

Gastwirth, F.C., and Cohen, M.L. (1970), Small Sample Behavior of Robust Linear Estimators of Location. *Journal of the American Statistical Association*, Vol. 65, p. 946.

Gross, A.M. (1973), Robust Confidence Intervals for the Location of Long Tailed Symmetric Distributions. Unpublished Ph.D. dissertation, Statistics Department, Princeton University.

Holland, P.W. (1975), A Variance Reduction Technique for Monte Carlo Studies of Robust Regression Confidence Intervals. NBER Working Paper, NBER Computer Research Center, Cambridge, MA.

McKinnon, W.J. (1964), Tables for Both the Sign Test and Distribution Free Confidence Intervals of the Median for Samples to 1,000. *Journal of the American Statistical Association*, Vol. 59, p. 935.

Nair, K.R. (1940), Table of Confidence Intervals for the Median in Samples from any Continuous Population. *Sankhya*, Vol. 4, p. 551.

Noether, G.E., (1949), Confidence Limits in the Nonparametric Case. *Journal of the American Statistical Association*, Vol. 44, p. 89.

Scheffe, H. (1943), Statistical Inference in the Nonparametric Case. *Annals of Mathematical Statistics*, Vol. 14, p. 305.

Tukey, J.W. (1970), *Exploratory Data Analysis*. (Limited preliminary edition), Addison-Wesley.

Weisberg, H.S. (1973), Contributions to Order Statistics. Unpublished Ph.D. dissertation, Statistics Department, Harvard University.