

## General Disclaimer

### One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.
- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.
- This document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

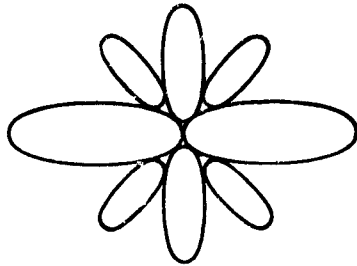
**EXACT INTERVALS AND TESTS FOR MEAN OF SYMMETRICAL POPULATION**

**WHEN ONE "SAMPLE" VALUE POSSIBLY AN OUTLIER**

by

**John E. Walsh**

**Technical Report No. 88  
Department of Statistics ONR Contract**



Reproduction in whole or in part is permitted  
for any purpose of the United States Government

This document has been approved for public release  
and sale; its distribution is unlimited.

**DEPARTMENT OF STATISTICS  
Southern Methodist University  
Dallas, Texas 75222**



FACILITY FORM 502

**N71-26091**  
ACCESSION NUMBER

PAGES: **118504**  
(NASA CR OR TMX OR AD NUMBER)

(THRU) **63**  
(CODE)

(CATEGORY)

**EXACT INTERVALS AND TESTS FOR MEAN OF SYMMETRICAL POPULATION**

**WHEN ONE "SAMPLE" VALUE POSSIBLY AN OUTLIER**

by

**John E. Walsh**

**Technical Report No. 88  
Department of Statistics ONR Contract**

**December 1, 1970**

**Research sponsored by the Office of Naval Research  
Contract N00014-68-A-0515  
Project NR 042-260**

**Reproduction in whole or in part is permitted  
for any purpose of the United States Government.**

**This document has been approved for public release  
and sale; its distribution is unlimited.**

**DEPARTMENT OF STATISTICS  
Southern Methodist University**

EXACT INTERVALS AND TESTS FOR MEAN OF SYMMETRICAL POPULATION  
WHEN ONE "SAMPLE" VALUE POSSIBLY AN OUTLIER

John E. Walsh  
Southern Methodist University\*

ABSTRACT

The (continuous) data are  $n$  observations that are believed to be a random sample from a symmetrical population. Confidence intervals and significance tests for the population mean are desired. There is, however, the possibility that either the smallest observation or the largest observation is an outlier. That is, the population providing this observation differs from the symmetrical population providing the other  $n - 1$  observations. If this occurs, intervals and tests are desired for the mean of the population providing the other  $n - 1$  observations. Some investigation difficulties can be overcome if intervals and tests can be developed that are simultaneously usable for all of these three situations (a confidence coefficient, or significance level, has the same value for all three situations). Two kinds of intervals and tests with this property are developed. These results always involve both the next to largest and next to smallest observations and should have at least moderately high efficiencies. Also, some extensions are considered, such as allowing each observation to be from a different population.

---

\* Research partially supported by Mobil Research and Development Corporation and by NASA Grant NGR 44-007-028. Also associated with ONR Contract N00014-68-A-0515.

INTRODUCTION AND DISCUSSION

The data are  $n$  independent observations that are continuous data and are believed to be a random sample from a symmetrical population. The order statistics of these observations are

$$X_1 < \dots < X_n.$$

Desired are confidence intervals and significance tests for the mean  $\mu$  of the population sampled. However, the experimental situation is such that  $X_1$  or  $X_n$ , but not both, could possibly be an outlier. More specifically, the population providing the observation that is  $X_1$ , or is  $X_n$ , differs from the symmetrical population that provides the other  $n - 1$  observations. Inequality of means is the difference of predominant interest but lack of symmetry for the population providing the outlier can also be of interest. When this outlier situation occurs, intervals and tests are desired for the mean  $\mu$  of the symmetrical population providing the other  $n - 1$  observations. The motivation for considering this outlier possibility could be due to almost anything (previous experience with data of this type, an examination of the observed values, a desire to protect against this possibility, etc.).

If  $X_1$  is an outlier,  $X_2, \dots, X_n$ , are the order statistics of a random sample of size  $n - 1$  from a symmetrical population. Here,  $X_2$  represents the smallest sample value, etc. If  $X_n$  is an outlier,  $X_1, \dots, X_{n-1}$  are the order statistics of a random sample of size  $n - 1$  from a symmetrical population.

Given a satisfactory procedure for deciding on the rejection of an outlier, one approach would be to first use this procedure to decide

which of the three situations (random sample,  $X_1$  an outlier,  $X_n$  an outlier) exists. A procedure for rejection of outlying observations for a general continuous symmetrical population is given in ref. 1. However, this procedure is not usable for a single outlier. Moreover, the decision reached might be erroneous even if a satisfactory procedure for rejection of an outlier were available.

A more desirable approach is to obtain intervals and tests that simultaneously are applicable to all three situations. More specifically, an interval has the same confidence coefficient, and a test has the same significance level, for the three situations. Two-sided intervals and tests with this property can be developed and two types are given. For brevity, only confidence intervals are considered. However, the corresponding tests can be obtained from these intervals and the usual manner.

The normal distributions are examples of symmetrical distributions. Thus, the results presented can be used to investigate  $\mu$  for the normality case. These results should have reasonably high efficiency for each of the three situations when the population providing  $n$ , or  $n - 1$ , sample values is normal. This efficiency property is indicated by some power computations given in ref. 2 for the case where  $X_2$  and  $X_{n-1}$  are used (appropriately) for a sample of size  $n$ . Of course, comparison with the corresponding best results separately for each of the three situations is somewhat unfair to the results given here. That is, the best results for a situation are only usable for that situation. A more meaningful comparison would be with the best results that are simultaneously usable for all three situations. Then, the results given here would seem to be highly efficient for almost any kind of underlying distribution.

The intervals and tests are of a symmetrical nature (same probability for each "tail") when the random sample situation occurs. The results for the outlier situations are two-sided but the probability for one tail does not even roughly equal the probability for the other tail. Thus, an unequal emphasis occurs for each outlier situation and should be taken into consideration when the results given here are candidates for use.

The next section contains a statement of the two types of confidence intervals and verification that a confidence coefficient has the same value for the three situations. An outline of some extensions, including use when each observation can be from a different population, is given in the final section.

#### RESULTS AND VERIFICATION

Direct use of material given in refs. 2 and 3 shows that, for a random sample of size  $n$  from a continuous symmetrical population,

$$P[(X_2 + X_{i+1})/2 \leq \mu \leq (X_{n-1} + X_{n-i})/2] = 1 - (n + 2 - i)2^{-(n - i)},$$

$$P\{\min[X_3, (X_2 + X_{i+1})/2] \leq \mu \leq \max[X_{n-2}, (X_{n-1} + X_{n-i})/2]\} \\ = 1 - [ni - (d + 1)(d - 2)/2]2^{-(n - 1)}.$$

These relationships define the two types of confidence intervals that are considered, and furnish their confidence coefficients ( $1 \leq i \leq n/2 - 1$ ).

Now, verification is given for the assertion that a confidence coefficient has the same value for the three situations. Proof for the

situation where  $X_n$  is an outlier is sufficient, since a completely analogous proof holds for the situation of  $X_1$  an outlier.

When  $X_n$  is an outlier, the sample size is  $n - 1$  and  $X_1, \dots, X_{n-1}$  are the sample order statistics. Use of material on page 160 of ref. 4, and some of the above results, for sample size  $n - 1$  shows that

$$P[(X_2 + X_{i+1})/2 > \mu] = (n + 1 - i)2^{-(n - i)},$$

$$P[(X_{n-1} + X_{n-i})/2 < \mu] = 2^{-(n-i)},$$

$$P\{\min[X_3, (X_2 + X_{i+1})/2] < \mu\} = [(n - 1)i - (d + 1)(d - 2)/2]2^{-(n - 1)},$$

$$P\{\max[X_{n-2}, (X_{n-1} + X_{n-i})/2] > \mu\} = i 2^{-(n - 1)}.$$

Summation of the two tail probabilities for each type and subtraction from unity verifies that a confidence coefficient for this outlier situation equals the corresponding confidence coefficient for the random sample situation.

#### COMMENTS ON EXTENSIONS

The results given are expressed in a way that is ordinarily used when the possibility of an outlier is considered. The requirements of random samples are, however, not needed. The two types of intervals and tests are usable under more general circumstances. That is, they are applicable when the conditions are such that the observations are independent and from continuous symmetrical populations that are believed to have a common mean  $\mu$ . However, either  $X_1$  or  $X_n$  could be an outlier. Here, "outlier" implies that the population providing this observation



has a mean different from the common mean  $\mu$  for the symmetrical populations providing the other  $n - 1$  observations, and/or that the outlier population is not symmetrical.

The mean might not exist for some populations. Then, the median can be investigated and  $\mu$  represents the central median (center median in interval of median values, when the median is not unique).

Continuous populations are not required if ties are resolved by use of an independent probability process (for example, see pages 139-140 of ref. 4). In any case, the true confidence coefficient is at least as large as the value for continuous populations.

Finally, other types of results likely could be developed that have the same confidence coefficient value for the three situations. The two types presented have the advantage of simplicity and of using both  $X_2$  and  $X_{n-1}$ .

REFERENCES

1. John E. Walsh, "Some nonparametric tests of whether the largest observations of a set are too large or too small," Annals of Mathematical Statistics, Vol. 21 (1950), pp. 583-592.
2. John Edward Walsh, Some Significance Tests for the Median Which Are Valid Under Very General Conditions. Unpublished Doctoral Thesis Princeton University, 1947, 87pp.
3. John E. Walsh, "Some significance tests for the median which are valid under very general conditions," Annals of Mathematical Statistics, Vol. 20 (1949), pp. 64-81.
4. John E. Walsh, Handbook of Nonparametric Statistics, D. Van Nostrand Co., Inc., Princeton, N. J., 1962, 575 pp.

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)  SOUTHERN METHODIST UNIVERSITY	2a. REPORT SECURITY CLASSIFICATION  UNCLASSIFIED
	2b. GROUP  UNCLASSIFIED

3. REPORT TITLE  
"Exact intervals and tests for mean of symmetrical population when one "sample" value possibly an outlier."

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)  
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)  
John E. Walsh

6. REPORT DATE December 1, 1970	7a. TOTAL NO. OF PAGES 7	7b. NO. OF REFS 4
------------------------------------	-----------------------------	----------------------

8a. CONTRACT OR GRANT NO. N00014-68-A-0515 b. PROJECT NO. NR 042-260 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S)  88
	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT  
This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY  Office of Naval Research
-------------------------	--

13. ABSTRACT  
The (continuous) data are n observations that are believed to be a random sample from a symmetrical population. Confidence intervals and significance tests for the population mean are desired. There is, however, the possibility that either the smallest observation or the largest observation is an outlier. That is, the population providing this observation differs from the symmetrical population providing the other n - 1 observations. If this occurs, intervals and tests are desired for the mean of the population providing the other n - 1 observations. Some investigation difficulties can be overcome if intervals and tests can be developed that are simultaneously usable for all of these three situations (a confidence coefficient, or significance level, has the same value for all three situations). Two kinds of intervals and tests with this property are developed. These results always involve both the next to largest and next to smallest observations and should have at least moderately high efficiencies. Also, some extensions are considered, such as allowing each observation to be from a different population.