EXACT INTERVALS AND TESTS FOR MEDIAN WHEN ONE "SAMPLE" VALUE POSSIBLY AN OUTLIER

by

John E. Walsh

Technical Report No. 87
Department of Statistics ONR Contract

December 18, 1970

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

DEPARTMENT OF STATISTICS
Southern Methodist University

INTRODUCTION AND DISCUSSION

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

$$x(1) < x(2) < \ldots < x(n).$$

Confidence intervals and significance tests are desired for the median θ (not necessarily unique) of the population sampled. However, the experimental situation is such that either x(1) or x(n) might possibly be an outlier. That is, the population yielding the observation that is x(1), or the observation x(n), and the population providing the other n - 1 observations do not have a common median. If this outlier situation should happen to exist, intervals and tests are desired for the median θ of the population yielding the n - 1 other observations. Incidentally, recognition of this outlier possibility could arise in any manner (examination of the observations, past experience with data from this source, the desire to be careful, etc.).

When x(1) is an outlier, x(2), ... , x(n) constitute a random sample of size n - 1. In this sample, x(2) is the smallest value, x(3) is the next to smallest value, etc. Likewise, x(1), ... , x(n - 1) provide a random sample of size n - 1 when x(n) is an outlier. In this sample, x(n - 1) is the largest value, x(n - 2) is the next to largest value, etc.

One approach to this investigation problem is to first develop a method for deciding which of the three situations (random sample, x(1) an outlier, x(n) an outlier) exists. Then, intervals and tests for θ

continuous populations that are believed to have a common median θ. Also, results can be obtained for cases where the data are not continuous.

VERIFICATION

Only the situation where x(1) is an outlier receives consideration. A similar method provides verification that (1) holds when x(n) is an outlier.

In general, the value of $P[x(i) \leq \theta \leq x(n+1-i)]$ can be expressed as unity minus

$$P[x(i) > \theta] + P[x(n+1-i) < \theta].$$

When x(1) is an outlier,

$$P[x(i) > \theta] = (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-1} \binom{n-1}{j},$$

$$P[x(n+1-i) < \theta] = (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-2} \binom{n-1}{j},$$

and their sum is

$$(\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-1} \left[\binom{n-1}{j} + \binom{n-1}{j-1}\right],$$

where $\binom{n-1}{-1}$ is zero. However, $\binom{n-1}{0} = \binom{n}{0}$ and

$$\binom{n-1}{j} + \binom{n-1}{j-1} = \binom{n}{j}$$

for $1 \leq j < i$. Thus, the value of $P[x(i) \leq \theta \leq x(n+1-i)]$ is

$$1 - (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-1} \binom{n}{j},$$

which is the value of (1). It is to be noticed that $P[x(i) > \theta]$ does not differ much from $P[x(n+1-i) < \theta]$ when i is of at least moderate size (ordinarily implies that n is at least moderately large).

EXTENSIONS

The preceding results are stated in the manner commonly used when cosidering the possibility of an outlier. However, the random sample requirements are not necessary. The intervals and tests apply, exactly, under more general conditions. Specifically they are usable when the circumstances are such that the observations are independent and from continuous populations that are believed to have a common median θ (not necessarily unique). However, either x(1) or x(n) might be an outlier, in the sense that the population yielding this observation has a median that is different from the common median θ for the populations yielding the other n - 1 observations.

The requirement of continuous populations is unnecessary if ties in observed values are resolved by randomization (all possibilities equally likely). Then, the confidence coefficients and significance levels are still exact. In any case,

$$P[x(i) \le \theta \le x(n+1-i)] \ge 1 - (\tfrac{1}{2})^{n-1} \sum_{j=1}^{i-1} \binom{n}{j}$$

for all three situations.

UNCLASSIFIED
Security Classification

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 median 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 18, 1970

7a. TOTAL NO. OF PAGES
5

7b. NO. OF REFS
0

8a. CONTRACT OR GRANT NO.
N00014-68-A-0515

b. PROJECT NO.
NR 042-260

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
87

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

Available are n observations (continuous data) that are believed to be a random sample. Desired are confidence intervals and significance tests for the population median. However, there is the possibility that either the largest or the smallest observation is an outlier. That is, the population yielding this observation differs from the population yielding the other n - 1 observations. If this happens, intervals and tests are desired for the median of the population yielding the n - 1 observations. Some analysis difficulties would be avoided if intervals and tests could be developed that simultaneously are applicable for all three of these situations. More specifically, a confidence coefficient, or significance level, has the same value for all three situations. It is found that two-sided intervals and tests based on two symmetrically located order statistics (not the largest and smallest) have this prpperty. Also, some extensions are considered wherein each observation can be from a different population.

UNCLASSIFIED
Security Classification