NONPARAMETRIC ESTIMATION OF MEAN AND VARIANCE WHEN A FEW "SAMPLE" VALUES POSSIBLY OUTLIERS

by

John E. Walsh

Technical Report No. 91
Department of Statistics ONR Contract

DEPARTMENT OF STATISTICS
Southern Methodist University
Dallas, Texas 75222

NONPARAMETRIC ESTIMATION OF MEAN AND VARIANCE WHEN A FEW "SAMPLE" VALUES POSSIBLY OUTLIERS

by

John E. Walsh

Technical Report No. 91
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

NONPARAMETRIC ESTIMATION OF MEAN AND VARIANCE WHEN A FEW "SAMPLE" VALUES POSSIBLY OUTLIERS

John E. Walsh

Southern Methodist University*

ABSTRACT

The data (continuous) are n independent observations that are believed to be a random sample. The possibility exists, however, that as many as J of the largest observations, and as many as K of the smallest observations, are outliers. That is, these observations are from populations that are different from the population yielding the other observations (which number at least n-J-K). The interest is in obtaining suitable estimates for the mean and variance of the population yielding the other observations. J and K are given and relatively small, with both $\leq 2n^{A}$, where A is specified and $\leq 1/4$. When the population yielding the other observations is continuous, has moments of all orders, and is well-behaved in some other ways, estimates are developed that are unbiased if terms of order $n^{-1+A+2\epsilon}$ are neglected. Here, ϵ can be arbitrarily small but is positive.

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

INTRODUCTION AND RESULTS

The data are n independent observations from continuous univariate populations. These observations are believed to be a random sample and estimates are desired for the population mean and variance. However, there is the possibility that as many as J of the largest observations and as many as K of the smallest observations are from populations that differ from the population yielding the other observations. Then, the interest is in obtaining suitable estimates for the mean μ and the variance σ^2 of the population yielding the random sample (of size at least n-J-K) that consists of the other observations. The values of J and K are given and relatively small. Specifically, $0 \le J, K \le 2n^A$, where A is given and such that $0 \le A \le 1/4$.

Let the order statistics of the n observations be denoted by

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

Then, x(1), ..., x(k) and x(n+1-j), ..., x(n) are from populations that differ from the population yielding x(k+1), ..., x(n-j), which constitute a random sample of size n-j-k. Here, j = 0 implies that none of the largest observations are from differing populations and k = 0 implies that none of the smallest observations are from differing populations. The values of j and k are unknown but satisfy $j \le J$ and $k \le K$.

The properties stated for the estimates presented do not hold in general. These estimates are not applicable unless n is at least moderately large and the population yielding the random sample of size n satisfies some conditions (at least approximately). Besides being continuous, this population should have finite moments of all orders and should have a density function that is analytic and nonzero throughout the range of possible values. A more exact statement of these conditions is given in the Derivations section.

The estimates could be stated in many ways. The statement given here uses all of $x(k+1),\ldots,x(n-j)$ with equal weighting. These are the only observations that are known to be from the population with mean μ and variance σ^2.

The estimate of μ is denoted by $\bar{x}(J,K)$ and the estimate of σ^2 is $S(J,K)$, where $\bar{x}(J,K)$ equals

$$(n-J-K)^{-1}[x(K+1) + x(K+2) + \ldots + x(n-J)]$$

and $S(J,K)$ equals

$$(n-J-K-1)^{-1}[x(K+1)^2 + \ldots + x(n-J)^2] - [(n-J-K)/(n-J-K-1)]\bar{x}(J,K)^2 .$$

These estimates have the properties

$$E[\bar{x}(J,K)] = \mu + O(n^{-1+A+\epsilon}),$$

$$E[S(J,K)] = \sigma^2 + O(n^{-1+A+2\epsilon}),$$

$$\mathrm{Var}[\bar{x}(J,K)] = \sigma^2/n + o(n^{-1}),$$

$$\mathrm{Var}[S(J,K)] = O(n^{-1}),$$

where $\epsilon > 0$ is a fixed but arbitrarily small constant. It is to be remembered that 1/4 is the largest possible value for A.

The next, and final, section contains an outline of the derivations for the properties of $\bar{x}(J,K)$ and $S(J,K)$.

OUTLINE OF DERIVATIONS

The relationships occurring in the derivations are similar to those arising in ref. 1. For brevity, much of the verification is only outlined, with referral to ref. 1 for more details.

The basic approach is to state $\bar{x}(J,K)$ and $S(J,K)$ in terms of $x(k+1)$, $\ldots, x(n-j)$, which is a random sample from the population considered, plus additional terms. Then, expressions whose expectations are μ and σ^2, respectively, can be identified and the additional terms are shown to be unimportant for n sufficiently large.

Some notation is introduced first. The mean of the sample of size $n-j-k$ is denoted by $\bar{x}(j,k)$ and is obtained from the expression for $\bar{x}(J,K)$ by letting $J = j$ and $K = k$. The arithmetic average of the order statistics $x(k+1),\ldots,x(K),x(n-J+1),\ldots,x(n-j)$ is denoted by y and the arithmetic average of the squares of these order statistics is represented by Y^2.

Let $F(x)$ be the cumulative distribution function of the population yielding $x(k+1),\ldots,x(n-j)$, and let $X^{(t)}(z)$, for $t = 0,1,2,\ldots$, be defined by

$$F[X^{(0)}(z)], \qquad X^{(t)}(z) = d^tX^{(0)}(z)/dz^t.$$

The more exact conditions on $F(x)$ are: $X^{(0)}(z)$ can be expanded in Taylor series about each of the values $z = (k+1)/(n-j-k),\ldots,K/(n-j-k)$, $(n-J+1)/(n-j-k),\ldots,(n-j)/n-j-k)$ and, for each series, $\int_0^1 [X^{(0)}(z)]^b dz$ can be evaluated using term by term integration ($b=1,\ldots,4$). Also, the magnitude of $z^tX^{(t)}(z)$ is at most $O(1)$ with respect to n for these values of z, $(t=1,2,\ldots)$, and the $X^{(0)}(z)$ are at most $O(n^{\epsilon})$, where $\epsilon > 0$ is arbitrarily small but a fixed constant. For $t = 2,3,\ldots$, the magnitude of $z^{t}X^{(t)}(z)$ is at most $o(1)$ for these values of z.

These conditions (taken from ref. 1) are not very restrictive for practical situations involving continuous populations. The first part justifies some expansions that are used. The magnitude relationships for the $X^{(0)}(z)$ are motivated by the consideration that this is the case when all the population moments exist. The relationships involving the $X^{(t)}(z)$ for $t \geq 1$ hold for nearly all continuous populations of practical interest.

The expectation of $\bar{x}(J,K)$ is considered first. The value of $\bar{x}(J,K)$ can be expressed as

$$[(n-j-k)/(n-J-K)]\bar{x}(j,k) + [(J+K-j-k)/(n-J-K)]y$$

Thus,

$$E[\bar{x}(J,K)] = \mu + O(n^{-1+A+\epsilon}),$$

since

$$E[\bar{x}(j,k)] = \mu, \qquad E(y) = O[(n-j-k)^{\epsilon}]$$

and j,k,J,K are $O(n^{A})$.

Next, consider the variance of $\bar{x}(J,K)$. By a method very similar to that used in ref. 1 (for the variance of m_x considered there), the variance of $\bar{x}(J,K)$ is found to be $\sigma^2/n + o(n^{-1})$. The principal use of this result is in evaluation of the expectation of $S(J,K)$. Another result for this purpose is

$$E(Z^2) = \mathrm{Var}(Z) + [E(Z)]^2,$$

which applies, in particular, when Z is an order statistic. From the stated conditions, and material in ref. 1,

$$E(Z^2) = O[(n-j-k)^{2\epsilon}]$$

when Z is any of $x(k+1),\ldots,x(K),x(n-J+1),\ldots,x(n-j)$.

Now, consider the expectation of $S(J,K)$. The value of $S(J,K)$ can be expressed as

$$[(n-j-k-1)/(n-J-K-1)](n-j-k-1)^{-1}[x(k+1)^2 + \ldots + x(n-j)^2]$$

$$- [(J+K-j-K)/(n-J-K-1)]Y^2$$

$$- [(n-J-K)/(n-J-K-1)]\bar{x}(J,K)^2.$$

Thus, $E[S(J,K)]$ equals

$$[(n-j-k-1)/(n-J-K-1)](\sigma^2+\mu^2) - (J+K-j-k)(n-J-K-1)^{-1}O[(n-j-k)^{2\epsilon}]$$

$$- [(n-J-K)/(n-J-K-1)][\sigma^2/n+o(n^{-1}) + \mu^2 + O(n^{-1+A+\epsilon})$$

$$= \sigma^2 + O(n^{-1+A+2\epsilon}).$$

The fact that $Var[S(J,K)]$ is $O(n^{-1})$ is verified by a method very similar to that used in ref. 1 (for the variance of s_x^2 considered there).

REFERENCE

1. John E. Walsh, "Nonparametric mean and variance estimation from truncated data," Skandinavisk Aktuarietidskrift, Vol 41 (1958), pp. 125-130.

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

"Nonparametric estimation of mean and variance when a few "sample" values possibly outliers"

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	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 18, 1970	6	1

8a. CONTRACT OR GRANT NO.

N00014-68-A-0515

b. PROJECT NO.

NR 042-260

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

91

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 data (continuous) are n independent observations that are believed to be a random sample. The possibility exists, however, that as many as J of the largest observations, and as many as K of the smallest observations, are outliers. That is, these observations are from populations that are different from the population yielding the other observations (which number at least $n - J - K$). The interest is in obtaining suitable estimates for the mean and variance of the population yielding the other observations. J and K are given and relatively small, with both $\leq 2n^A$, where A is specified and $\leq 1/4$. When the population yielding the other observations is continuous, has moments of all orders, and is well-behaved in some other ways, estimates are developed that are unbiased if terms of order $n^{-1+A+2\varepsilon}$ are neglected. Here, ε can be arbitrarily small but is positive.

DD FORM 1 NOV 65 1473 (PAGE 1)

UNCLASSIFIED