THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

SAMPLE SIZES FOR APPROXIMATE INDEPENDENCE BETWEEN SAMPLE MEDIAN AND LARGEST (OR SMALLEST) ORDER STATISTIC

by

John E. Walsh

Technical Report No. 7
Department of Statistics THEMIS Contract

Department of Statistics
Southern Methodist University
Dallas, Texas 75222

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

SAMPLE SIZES FOR APPROXIMATE INDEPENDENCE BETWEEN SAMPLE MEDIAN AND LARGEST (OR SMALLEST) ORDER STATISTIC

by

John E. Walsh

Technical Report No. 7
Department of Statistics THEMIS Contract

September 11, 1968

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

DEPARTMENT OF STATISTICS
Southern Methodist University

SAMPLE SIZES FOR APPROXIMATE INDEPENDENCE BETWEEN SAMPLE MEDIAN AND LARGEST (OR SMALLEST) ORDER STATISTIC

by

John E. Walsh
Southern Methodist University*
Dallas, Texas

Let $X_1 \leq \ldots \leq X_{2n+1}$ be the order statistics for a random sample of size $2n + 1$. Asymptotically, X_{n+1} and X_{2n+1} are independent. That is, the maximum of the differences between $P(X_{n+1} \leq x_{n+1}, X_{2n+1} \leq x_{2n+1})$ and the corresponding values assuming independence tends to zero as $n \to \infty$. A minimum sample size is (approximately) determined which assures that the maximum difference is at most a stated amount. This minimum sample size is the smallest allowable for continuous populations but smaller sample sizes could possibly be usable for discontinuous cases. Likewise, X_1 is asymptotically independent of X_{n+1} and this same minimum sample size is applicable for the stated maximum difference. The minimum sample size is finite for all nonzero maximum differences but is very large if the maximum difference is much smaller than .005.

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

INTRODUCTION AND RESULTS

In general, the largest order statistic of a sample, also the smallest order statistic, are asymptotically independent of the sample median. That is, the maximum of the nonnegative difference

$$P(X_{n+1} \le x_{n+1}, X_{2n+1} \le x_{2n+1}) - P(X_{n+1} \le x_{n+1})P(X_{2n+1} \le x_{2n+1}), \quad (1)$$

over x_{n+1} and x_{2n+1}, tends to zero as $2n+1 \to \infty$, where $X_1 \le \ldots \le X_{2n+1}$ are the order statistics for a sample of size $2n+1$ from any possible population. Also, the maximum of the nonnegative difference

$$P(X_1 \le x_1, X_{n+1} \le x_{n+1}) - P(X_1 \le x_1)P(X_{n+1} \le x_{n+1}), \quad (2)$$

over x_1 and x_{n+1}, tends to zero as $2n+1 \to \infty$.

The maximum of the differences between the true joint probabilities and the corresponding values assuming independence is a measure of the level of independence. Minimum sample sizes, for assuring that the maximum difference of (1) is at most stated amounts, are approximately determined. These minimum sample sizes also assure that the maximum difference of (2) is at most the stated amounts. Let ϵ be the specified maximum difference. The value of (1), also that of (2), is (approximately) at most ϵ when

$$2n+1 \ge -1 + e^{-2}/2\pi\epsilon 2 \doteq -1 + .0215/\epsilon^2, \qquad (\epsilon \le .02).$$

For example, let $\epsilon = .005$. Then the sample size is at least 859.

Although only the situation of odd sample sizes is explicitly considered, these results should also be applicable for even sample sizes (say, with $\epsilon \le .015$). Then the sample median is the arithmetic average of the two central order statistics.

The lower bounds for sample sizes are developed under the assumption that the individual probability expressions occurring in

(1) and (2) can take all values between zero and one (continuous case). When this does not happen, sample sizes less than those (approximately) dictated by the inequality could possibly occur. That is, the probability expressions may not be able to take on the values that maximize the lower bound for the sample size.

The next and final section contains an outline of the derivations for the results stated above. Only the case of X_{n+1} and X_{2n+1} is considered. Derivations for the case of X_1 and X_{n+1} are of an analogous nature.

DERIVATIONS

The meaningful values of x_{n+1} and x_{2n+1} are those that correspond to population percentiles of $P(X_{n+1} \le x_{n+1})$ and of $P(X_{2n+1} \le x_{2n+1})$. Let the values of $P(X_{2n+1} \le x_{2n+1})$ be represented by e^{-b}, and $(0 \le b < \infty)$, while $P(X_{n+1} \le x_{n+1}) = \alpha$, $(0 < \alpha < 1)$. All values of b and all values of α are possible (continuous case). Then, the difference (1) can be expressed as

$$P(X_{2n+1} \le x_{2n+1})[P(X_{n+1} \le x_{n+1} | X_{2n+1} \le x_{2n+1}) - P(X_{n+1} \le x_{n+1})]$$
$$= e^{-b}[P(X_{n+1} \le x_{n+1} | X_{2n+1} \le x_{2n+1}) - \alpha].$$

The initial problem is to evaluate $P(X_{n+1} \le x_{n+1} | X_{2n+1} \le x_{2n+1})$ in terms of n, b, and α. Then, the difference (1) is set equal to ϵ. Finally, an expression for the value of n that yields ϵ is maximized with respect to α and b (actually, a monotonic function of α is considered).

Let F denote the probability that a sample value is at most equal to x_{n+1}. Then the conditional probability is $F' = Fe^{-b/(2n+1)}$

that a sample value is at most x_{n+1} when it is given that $X_{2n+1} \le x_{2n+1}$. Evaluation of F for any stated value of α is considered next.

Using the material of (Feller, 1945), $P(X_{n+1} \le x_{n+1})$ = (probability number of observations with values $\le x_{n+1}$ is at least n+1) can be expressed in the form

$$\exp\{5[1 - F(1 - F)]/2(n + 1)F(1 - F)\} \tag{3}$$

$$X\left(1 - \Phi\{[(n + 1)(1 - 2F) + a(F,n)]/[2(n + 1)F(1 - F)]^{1/2}\}\right),$$

where $\Phi\{x\}$ is the standardized normal cumulative distribution function and $a(F,n)$ is $O(1)$ with respect to n. Also, $a(F + O(n^{-1}), n)$ equals $a(F,n) + O(n^{-1})$. The expression (3) is a monotonically increasing function of F.

If the difference (1) is to be ϵ, it is seen that $\alpha \le 1 - \epsilon$ and $b \le -\log_e \epsilon$ in all cases. Also $P(X_{n+1} \le x_{n+1} | X_{2n+1} \le x_{2n+1})$ is at least ϵ. These relations imply that b is not large and that F is nontrivially bounded away from zero and unity. Thus, expansions can be made in terms of b divided by functions of n that tend to ∞ as $n \to \infty$. Also, examination shows that $F = 1/2 + O(n^{-1/2})$, so that $F(1 - F) = 1/4 + O(n^{-1})$. This implies that

$$e^{15/2(n+1)}\left(1 - \Phi[(n+1)(1-2F) + a(F,n)]/[2(n+1)F(1-F)]^{1/2}\right) + O(n^{-2})$$

equals (3).

Let $\beta = \alpha e^{-15/2(n+1)}$. Then, the value of $P(X_{n+1} \le x_{n+1}) = \alpha + O(n^{-2})$ when

$$[(n + 1)(1 - 2F) + a(F,n)]/2(n + 1)F(1 - F)]^{1/2} = K_\beta.$$

Squaring both sides and solving the quadratic in F for the appropriate root yields

$$F = 1/2 + a(F,n)/[2(n+1) + K_\beta^2] - (1/2)K_\beta[2(n+1) + K_\beta^2]^{-1/2}$$

$$= 1/2 + a(F',n)/2(n+1) - (1/2)K_\beta[2(n+1)]^{-1/2}[1 - K_\beta^2/4(n+1)] + O(n^{-2})$$

as an (implicit) expression for F.

Now, consider $P(X_{n+1} \le x_{n+1} | X_{2n+1} \le x_{2n+1})$. This can be expressed as

$$\exp\{5[1 - F'(1 - F')]/2(n+1)F'(1 - F')\}$$

$$X\left(1 - \Phi\{[(n+1)(1 - 2F') + a(F',n)]/[2(n+1)F'(1 - F')]^{1/2}\}\right)$$

$$= e^{15/2(n+1)}\left(1 - \Phi\left\{K_\beta - K_\beta^3/4(n+1) - b[2(n+1)]^{-1/2} + bK_\beta/2(n+1) + O(n^{-3/2})\right\}\right) + O(n^{-2}).$$

since $F'(1 - F') = 1/4 + O(n^{-1})$ and

$$[(n+1)(1 - 2F') + a(F',n)]/[2(n+1)F'(1 - F')]^{1/2}$$

$$= K_\beta - K_\beta^3/4(n+1) - b[2(n+1)]^{-1/2} + bK_\beta/2(n+1) + O(n^{-3/2}),$$

when the substitution

$$1/2 + a(F',n)/2(n+1) - (1/2)K_\beta[2(n+1)]^{-1/2}[1 - K_\beta^2/4(n+1)] + O(n^{-2})$$

is made for F in $F' = Fe^{-b/(2n+1)}$.

Thus, the difference (1) can be expressed as e^{-b} times

$$e^{15/2(n+1)}\left(\Phi\{K_\beta\} - \Phi\{K_\beta - K_\beta^3/4(n+1) - b[2(n+1)]^{-1/2} + bK_\beta/2(n+1) + O(n^{-3/2})\}\right)$$

plus $O(n^{-2})$. This expression equals e^{-b} times

$$\{b[2(n+1)]^{-1/2} + K_\beta^3/4(n+1) - bK_\beta/2(n+1)\}\,(2\pi)^{-1/2}$$

$$X \exp\left[-(1/2)\left(K_\beta - c(\beta,b,n)\{b[2(n+1)]^{-1/2} + K_\beta^3/4(n+1) - bK_\beta/2(n+1)\}\right)^2\right]$$

plus $O(n^{-3/2})$, where $0 \le c(\beta,b,n) \le 1$. Set this expression equal to ϵ. Then,

$$[2(n+1)]^{1/2} = e^{-b}\{b + K_\beta^3[8(n+1)]^{-1/2} - bK_\beta[2(n+1)]^{-1/2}\}(2\pi)^{-1/2}\,\epsilon^{-1}$$

$$\times \exp\left[-(1/2)\left(K_\beta - c(\beta,b,n)\{b[2(n+1)]^{-1/2} + K_\beta^3/4(n+1) - bK_\beta/2(n+1)\}\right)^2\right]$$

plus $O(n^{-1})$. The righthand side is maximum with respect to K_β when K_β is of the form $bc(\beta,b,n)/[2(n+1)]^{1/2} + O(n^{-1})$. With this substitution,

$$[2(n+1)]^{1/2} + O(n^{-1}) = be^{-b}/\epsilon(2\pi)^{1/2}.$$

The righthand side is maximum with respect to b when b = 1. Making this substitution and squaring both sides,

$$2(n+1) + O(n^{-1/2}) = e^{-2}/2\pi\epsilon^2 \doteq .0215/\epsilon^2$$

is enough to assure a difference of at most ϵ in all cases. The $O(n^{-1/2})$ term should be unimportant when $(2n+1)^{1/2}$ is greater than, say, seven. This is the case when $\epsilon \le .02$.

REFERENCE

Feller, W., "On the normal approximation to the binomial distribution," Annals of Math. Stat., Vol. 16, 1945, pp. 319-329.

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)

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

ORT TITLE

Sample Sizes for Approximate Independence Between Sample Median and Largest (or Smallest) Order Statistic

CRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

HOR(S) *(First name, middle initial, last name)*

John E. Walsh

ORT DATE September 11, 1968	7a. TOTAL NO. OF PAGES 7	7b. NO. OF REFS 1
NTRACT OR GRANT NO. N00014-68-A-0515 OJECT NO. NR 042-260	9a. ORIGINATOR'S REPORT NUMBER(S) 7. 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

TRIBUTION STATEMENT

No limitations

PLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

TRACT

Let $X_1 \leq \cdots \leq X_{2n+1}$ be the order statistics for a random sample of size $2n + 1$. Asymptotically, X_{n+1} and X_{2n+1} are independent. That is, the maximum of the differences between $P(X_{n+1} \leq x_{n+1}, X_{2n+1} \leq x_{2n+1})$ and the corresponding values assuming independence tends to zero as $n \to \infty$. A minimum sample size is (approximately) determined which assures that the maximum difference is at most a stated amount. This minimum sample size is the smallest allowable for continuous populations but smaller sample sizes could possibly be usable for discontinuous cases. Likewise, X_1 is asymptotically independent of X_{n+1} and this same minimum sample size is applicable for the stated maximum difference. The minimum sample size is finite for all nonzero maximum differences but is very large if the maximum difference is much smaller than .005.

FORM NOV 65 1473

UNCLASSIFIED
Security Classification