THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

APPROXIMATE JOINT PROBABILITIES FOR LARGEST AND SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

by

John E. Walsh

Technical Report No. 8
Department of Statistics THEMIS Contract

Department of Statistics
Southern Methodist University
Dallas, Texas 75222

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

APPROXIMATE JOINT PROBABILITIES FOR LARGEST AND SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

by

John E. Walsh

Technical Report No. 8
Department of Statistics THEMIS Contract

September 13, 1968

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

DEPARTMENT OF STATISTICS
Southern Methodist University

APPROXIMATE JOINT PROBABILITIES FOR LARGEST AND SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

John E. Walsh

Southern Methodist University*

ABSTRACT

Let X_n and X_1 be the largest and smallest, respectively, of a set of n independent observations. Also, let $\overline{F}(x;n)$ be the arithmetic average of the cumulative distributions for the individual observations. Often, the interest is in $P(X_1 > x_1, X_n \leq x_n)$ for practical applications. An approximate expression, also sharp upper and lower bounds, are developed for $P(X_1 > x_1, X_n \leq x_n)$ in terms of n and $\overline{F}(x_n;n) - \overline{F}(x_1;n)$. These results are applicable for x_n and x_1 such that $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] < 1$. The approximate expression is reasonably accurate if $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \leq .25$ and has a relative error of less than one percent when $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \leq .17$; then, $P(X_1 > x_1, X_n \leq x_n)$ is at least .75, and at least .83, respectively. All $n \geq 1$ and all possible distributions for the individual observations can occur. For continuity in its tails, approximate two-sided tolerance intervals using X_n and X_1 are developed for $\overline{F}(x;n)$. Approximate joint confidence regions and tests are obtained for an extreme upper and an extreme lower percentage point of $\overline{F}(x;n)$. Also, tests of $\overline{F}(x;n) \equiv F_0(x)$, completely specified, are developed for x in the tails.

* Research partially supported by NASA Grant NGR 44-007-028

Also associated with ONR Contract N00014-68-A-0515.

INTRODUCTION AND DISCUSSION

Consider a set of n independent observations and let X_n and X_1 denote the largest and smallest values, respectively. Often these is interest in whether X_n is unusually large and/or X_1 is unusually small. This has probability

$$1 - P(X_1 > x_1, X_n \le x_n),$$

where x_n is considered to be unusually large for X_n, and x_1 is unusually small for X_1. Quite accurate approximate probability expressions can be developed for this relation when x_n is sufficiently large and x_1 is sufficiently small. All $n \ge 1$ are considered and the individual observations can have any distributions.

Explicitly, an approximate expression is developed for $P(X_1 > x_1, X_n \le x_n)$ that is very accurate if $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \le .15$, where $\overline{F}(x;n)$ is the arithmetic average of the cumulative distribution functions(cdf's) for the separate observations. This expression is a function of

$n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)]$.

Sharp upper and lower bounds are developed for $P(X_1 > x_1, X_n \le x_n)$ in terms of n and $\overline{F}(x_n;n) - \overline{F}(x_1;n)$. The lower bound depends only on $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)]$ and for the situations of primary interest, this is approximately the case for the upper bound. These bounds apply only for $x_n$and x_1 such that $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] < 1$. They are very far apart for $n[1 - \overline{F}(x_n) - \overline{F}(x_1)] \ge .75$ but moderately close if

$n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \leq .5$ and very close when this quantity is at most .15. The approximate expression is about halfway between the bounds for the situations of principal interest.

As already indicated (Walsh, 1959,1964) the cdf $\overline{F}(x;n)$ occupies a central role for investigations using order statistics of sets of independent observations. Some procedures are given for investigating the tails of $\overline{F}(x;n)$.

Two-sided tolerance intervals for $\overline{F}(x;n)$ are obtained using X_n and X_1. Continuity of $\overline{F}(x;n)$ in the tails is assumed.

An extreme upper and an extreme lower percentage point of $\overline{F}(x;n)$ can be simultaneously investigated using X_n and X_1. This permits the "spread" of $\overline{F}(x;n)$ to be investigated. Confidence regions consisting of two simultaneous one-sided intervals (one for each percentage point) are easily developed when $\overline{F}(x;n)$ is continuous at the percentiles considered. Tests for simultaneously investigating specified values of these two percentiles are obtained from the confidence regions in the usual way. These regions and tests have rather accurate probability levels if the two percentiles correspond to values of x_n and x_1 such that $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \leq .2$. Even for such extreme percentiles, the probability levels are only bounded (instead of rather accurately determined) if $\overline{F}(x;n)$ is discontinuous at one or both of the percentiles considered.

Lastly, tests are developed for the null hypothesis that $\overline{F}(x;n) \doteqdot F_0(x)$, completely specified, in the tails. Here, the tails considered are for x_n and x_1 such that $\overline{F}(x_n;n) - \overline{F}(x_1;n) > 1 - 1/n$.

The next section contains a statement of the bounds, the approximate expression, and their derivation. Two-sided tolerance intervals are considered in the following section. The next to last section contains the joint confidence regions and tests for percentiles. The final section is concerned with tests of whether $\bar{F}(x;n)$ has a completely specified form in its tails.

BOUNDS AND APPROXIMATE EXPRESSION

When $n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)] < 1$, sharp upper and lower bounds for $P(X_1 > x_1, X_n \le x_n)]$ are provided by

$$1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]$$
$$\le P(X_1 > x_1, X_n \le x_n) \le [\bar{F}(x_n;n) - \bar{F}(x_1;n)]^n.$$

When $n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)] \le .25$, the value of $[\bar{F}(x_n;n) - \bar{F}(x_1;n)]^n$ approximately equals

$$1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)] + \frac{1}{2}\{n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]\}^2$$

and is less than this value. In these ranges of x_n and x_1, this expression can be used as an almost sharp upper bound. The arithmetic average of the lower bound and this upper bound is

$$1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)] + \frac{1}{4}\{n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]\}^2.$$

which is the approximate expression for $P(X_1 > x_1, X_n \le x_n)$.

Now consider derivation of the sharp bounds. Alternate proofs could be based on (Hoeffding, 1956). Let $F_i(x)$ denote the cdf for the i-th observation $(i = 1, \ldots, n)$. Then, $P(X_1 > x_1, X_n \le x_n)$ equals

$$\prod_{i=1}^{n} [F_i(x_n) - F_i(x_1)]$$

$$= \exp\left(\sum_{i=1}^{n} \log_e\{1 - [1 - F_i(x_n) + F_i(x_1)]\}\right)$$

$$= \exp\left\{-\sum_{i=1}^{n} \sum_{j=1}^{\infty} [1 - F_i(x_n) + F_i(x_1)]^j/j\right\}$$

$$= \exp\left\{- \sum_{j=1}^{\infty} j^{-1} \sum_{k=0}^{j} \binom{j}{k}[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]^{j-k}\right.$$

$$\left.\times \sum_{i=1}^{n} [\bar{F}(x_n;n) - \bar{F}(x_1;n) - F_i(x_n) + F_i(x_1)]^k\right\}$$

For $k \geq 2$ and $n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)] < 1$,

$$\sum_{i=1}^{n}[\bar{F}(x_n;n) - \bar{F}(x_1;n) - F_i(x_n) + F_i(x_1)]^k$$

is largest when all but one of the $F_i(x_n) - F_i(x_1)$ are unity and the remaining one is such that their arithmetic average is $\bar{F}(x_n;n) - \bar{F}(x_1;n)$. Then, the remaining $F_i(x_n) - F_i(x_1)$ equals

$$1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]$$

and

$$\sum_{i=1}^{n} [\bar{F}(x_n;n) - \bar{F}(x_1;n) - F_i(x_n) + F_i(x_1)]^k$$

$$= [(n - 1)^k + (- 1)^k(n - 1)][1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]^k.$$

Thus, since n^j equals

$$[(n - 1) + 1]^j + (n - 1)(1 - 1)^j = \sum_{k=0}^{j}\binom{j}{k}[(n - 1)^k + (-1)^k(n - 1)],$$

and $1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)$ to the k-th power is multiplied by this quantity to the (j-k)-th power, $P(X_1 > x_1, X_n \le x_n)$ is at least equal to

$$\exp\left\{-\sum_{j=1}^{\infty} j^{-1}\Big(n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]\Big)^j\right\}$$

$$=\exp\{\log_e(1 - n[1 - \bar{F}(X_n;n) + \bar{F}(x_1;n)])\}$$

$$= 1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)].$$

with equality possible.

The sharp upper bound is obtained from the relation that the geometric mean is at most equal to the arithmetic mean. Thus, raising both means to the n-th power,

$$\prod_{i=1}^{n}[F_i(x_n) - F_i(x_1)] \le [\bar{F}(x_n;n) - \bar{F}(x_1;n)]^n,$$

with equality possible.

TWO-SIDED TOLERANCE INTERVALS

The two-sided tolerance intervals considered for $\bar{F}(x;n)$ are of the form (X_1,X_n) and $\bar{F}(x;n)$ is required to be continuous in its tails.

The probability included in the random interval (X_1,X_n) equals $\bar{F}(X_n;n) - \bar{F}(x_1;n)$. Thus, the probability that (X_1,X_n) covers at least 100(1-p) percent of the probability for $\bar{F}(x;n)$ is

$$1 - P[\bar{F}(X_n;n) - \bar{F}(X_1;n) \le 1 - p] \qquad (1)$$

$$= 1 - \int_0^p P[p - p' < F(X_1) \le p - p' + dp',\ F(X_n) \le 1 - p']$$

where integration is with respect to dp' and terms of order $(dp')^2$ are neglected in the expansion of the probability within the integral.

Let $P(X_1 > x_1, X_n \leq x_n)$ be exactly represented as

$$1 - n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]$$

$$+ (1/2)\gamma(x_1,x_n)\{n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n)]\}^2$$

for $n[1 - \bar{F}(x_n;n) + \bar{F}(x_1;n) < 1$, where $\gamma(x_1,x_n)$ is continuous with zero and unity as bounds. Then, neglecting terms of order $(dp')^2$,

$$P(p - p' < X_1 \leq p - p' + dp', X_n \leq p')$$

$$= n[1 - \gamma(p - p', 1 - p')np]dp'.$$

Thus, the value of (1) is at least $1 - np$ and at most $1 - np(1 - np)$, where $p < 1/n$.

INVESTIGATION OF EXTREME PERCENTILES

The 100p percent point of $\bar{F}(x;n)$ is denoted by $\theta(p)$. Percentage points $\theta(p_n)$ and $\theta(p_1)$ are simultaneously investigated, where $1 - p_n + p_1 < 1/n$ and cases where $1 - p_n + p_1 \leq .2/n$ are of primary interest.

First, consider the case where $\bar{F}(x;n)$ is continuous (or very nearly so) at both $\theta(p_n)$ and $\theta(p_1)$. For $1 - p_n + p_1 \leq .2/n$,

$$P[X_1 > \theta(p_1), X_n \leq \theta(p_n)] \doteq [1 - n(1 - p_n + p_1)/2]^2.$$

Thus, the confidence region consisting of the random interval $(-\infty,X_1)$ for $\theta(p_1)$ and the random interval (X_n,∞) for $\theta(p_n)$ has a confidence coefficient of approximately $[1 -n(1 - p_n + p_1)/2]^2$.

Confidence regions of this form can also be obtained when $1 - p_n + p_1 > .2/n$ but their confidence coefficients are not as closely determined. The confidence coefficient bounds are only moderately close together for

$$.2/n < 1 - p_n + p_1 \leq .5/n$$

and are rather far apart in other cases.

When $\bar{F}(x;n)$ is continuous at $\theta(p_1)$ but not $\theta(p_n)$, the probability is at least that for continuity at $\theta(p_n)$. If the intervals are those in $P[X_1 > \theta(p_1), X_n < \theta(p_n)]$, the probability is at most that for continuity at $\theta(p_n)$. If $\bar{F}(x;n)$ is continuous at $\theta(p_n)$ but not at $\theta(p_1)$, the probability is at most that for continuity at $\theta(p_1)$. When $\bar{F}(x;n)$ is discontinuous at both $\theta(p_1)$ and $\theta(p_n)$, the value of $P[X_1 > \theta(p_1), X_n < \theta(p_n)]$ is at most that for continuity but a bound is not determined for $P[X_1 > \theta(p_1), X_n \leq \theta(p_n)]$.

HYPOTHESIS OF SPECIFIED DISTRIBUTION

Let $F_o(x)$ be the null form, completely specified, for $\bar{F}(x;n)$ in its tails. Also, only x_n and x_1 such that $1 - \bar{F}(x_n;n) + \bar{F}(x_1;n) \leq .2/n$ are considered for investigation. Then, $P(X_1 > x_1;X_n \leq x_n)$ approximately equals

$$\{1 - n[1 - F_o(x_n) + F_o(x_1)]/2\}^2$$

under the null hypothesis. This permits many kinds of investigations of the tails, although only one type is considered here.

Suppose that there is interest in disagreement between $\bar{F}(x;n)$ and $F_o(x)$ for the interval of x_n values such that $p_n^{(1)} < F_o(x_n) \leq p_n^{(2)}$ and also the x_1 such that $p_1^{(1)} < F_o(x_1) \leq p_1^{(2)}$. Here, $p_n^{(1)}$, $p_n^{(2)}$, $p_1^{(1)}$, $p_n^{(1)}$ are attainable values for $F_o(x)$ and, under the null hypothesis.

$$1 - P[p_1^{(1)} < F_o(X_1) \leq p_1^{(2)}, p_n^{(1)} < F_o(X_n) \leq p_n^{(2)}]$$

$$= 1 - P[\theta(p_1^{(1)}) < X_1 \leq \theta(p_1^{(2)}), \theta(p_n^{(1)}) < X_n \leq \theta(p_n^{(2)})]$$

has a small value that is suitable for a significance level. Then, the test that rejects $\bar{F}(x;n) \equiv F_o(x)$ in the tails if and only if one or both of $p_1^{(1)} < F_o(X_1) \leq p_1^{(2)}$ and $p_n^{(1)} < F_o(X_n) \leq p_n^{(2)}$ are not satisfied has this approximate significance level (for which upper and lower bounds can be determined).

One possible choice is $p_1^{(1)} = 0$ and/or $p_n^{(2)} = 1$ (one-sided or partially one-sided cases). Also, the critical region could be based on two or more (disjoint) intervals for one or both of $F_o(x_n)$ and $F_o(X_1)$.

REFERENCES

Hoeffding, Wassily, "On the distribution of the number of successes in independent trials," Annals of Math. Stat., Vol. 27, 1956, pp. 713-721.

Walsh, John E., "Definition and use of generalized percentage points," Sankhyā, Vol. 21, 1959, pp. 281-288.

Walsh, John E., "Approximate distribution of extremes for nonsample cases," Jour. Amer. Stat. Assoc., Vol. 59, 1964, pp. 429-436.

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

Approximate Joint Probabilities for Largest and Smallest
of a Set of Independent Observations

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 September 13, 1968	7a. TOTAL NO. OF PAGES 9	7b. NO. OF REFS 3
8a. CONTRACT OR GRANT NO. N00014-68-A-0515 b. PROJECT NO. NR 042-260 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 8. 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

10. DISTRIBUTION STATEMENT

No limitations

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

13. ABSTRACT

Let X_n and X_1 be the largest and smallest, respectively, of a set of n independent observations. Also, let $\overline{F}(x;n)$ be the arithmetic average of the cumulative distributions for the individual observations. Often, the interest is in $P(X_1 > x_1, X_n \le x_n)$ for practical applications. An approximate expression, also sharp upper and lower bounds, are developed for $P(X_1 > x_1, X_n \le x_n)$ in terms of n and $\overline{F}(x_n;n) - \overline{F}(x_1;n)$. These results are applicable for x_n and x_1 such that $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] < 1$. The approximate expression is reasonably accurate if $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \le .25$ and has a relative error of less than one percent when $n[1 - \overline{F}(x_n;n) + \overline{F}(x_1;n)] \le .17$; then, $P(X_1 > x_1, X_n \le x_n)$ is at least .75, and at least .83, respectively. All $n \ge 1$ and all possible distributions for the individual observations can occur. For continuity in its tails, approximate two-sided tolerance intervals using X_n and X_1 are developed for $\overline{F}(x;n)$. Approximate joint confidence regions and tests are obtained for an extreme upper and an extreme lower percentage point of $F(x;n)$ Also, tests of $\overline{F}(x;n) \equiv F_0(x)$, completely specified, are developed for x in the tails.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification