THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

APPROXIMATE DISTRIBUTIONS FOR LARGEST AND FOR SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

by

John E. Walsh

Technical Report No. 10
Department of Statistics THEMIS Contract

Department of Statistics
Southern Methodist University
Dallas, Texas 75222

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

APPROXIMATE DISTRIBUTIONS FOR LARGEST AND FOR SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

by

John E. Walsh

Technical Report No. 10
Department of Statistics THEMIS Contract

September 18, 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 DISTRIBUTIONS FOR LARGEST AND FOR SMALLEST OF A SET OF INDEPENDENT OBSERVATIONS

John E. Walsh

Southern Methodist University*

ABSTRACT

Consider n independent univariate observations with possibly different distributions. Let X_n and X_1 denote the largest and smallest observations, respectively, while $\overline{F}(x;n)$ is the arithmetic average of the cumulative distribution functions for the individual observations. Approximate expressions, also sharp upper and lower bounds, are developed for $P(X_n \le x)$ and $P(X_1 \le x)$ in terms of n and $\overline{F}(x;n)$. These results are applicable for $P(X_n \le x)$ when x is such that $n[1 - \overline{F}(x;n)] < 1$, and for $P(X_1 \le x)$ when $n\overline{F}(x;n) < 1$. The approximate expressions are reasonably near the bounds if $n[1 - \overline{F}(x;n)] \le .25$ and $n\overline{F}(x;n) \le .25$. Relative error is less than one percent if $n[1 - \overline{F}(x;n)] \le .17$ and $n\overline{F}(x;n) \le .17$; then, $P(X_n \le x) \ge .83$ and $P(X_1 \le x) \le .17$. All possible distributions and all $n \ge 1$ are allowable. Suitable estimation of $\overline{F}(x;n)$ provides estimates of $P(X_n \le x)$ and $P(X_1 \le x)$. In particular, for continuity in pertinent tail of $\overline{F}(x;n)$, asymptotic distributions are often obtained and the problem is simplified to estimating at most three parameters. Approximate confidence regions and tests are obtained for sufficiently extreme upper and lower percentage points of $\overline{F}(x;n)$. Also, for continuity in the pertinent tail, approximate tolerance intervals using X_n or X_1 are developed for $\overline{F}(x;n)$. In addition, tests of $\overline{F}(x;n) \equiv F_0(x)$, completely specified, are developed for x in the pertinent tail.

---

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

INTRODUCTION AND DISCUSSION

There is often interest in whether the largest observation of a set of n independent observations is unusually large, or the smallest observation is unusually small. Quite accurate approximate probability expressions can be developed for relations of this kind, even though the distributions for the individual observations can be arbitrarily different and all $n \geq 1$ are considered.

More specifically, let X_n and X_1 denote the largest and smallest observations, respectively. Approximate expressions are developed for $P(X_n \leq x)$ and $P(X_1 \leq x)$ that are very accurate if $1 - P(X_n \leq x) \leq .15$ and $P(X_1 \leq x) \leq .15$. The expression for $P(X_n \leq x)$ is a function of $n[1 - \overline{F}(x;n)]$, where $\overline{F}(x;n)$ is the arithmetic average of the cumulative distribution functions (cdf's) for the individual observations. The expression for $P(X_1 \leq x)$ is a function of $n\overline{F}(x;n)$.

Generally, sharp upper and lower bounds are developed for $P(X_n \leq x)$ in terms of n and $\overline{F}(x;n)$. The lower bound depends only on $n[1 - \overline{F}(x;n)]$ and, for the situations of principal interest, this is approximately the case for the upper bound. The approximate expression is about halfway between these bounds for the situations of principal interest. These results are applicable only for x such that $n[1 - \overline{F}(x;n)] < 1$. The bounds are quite far apart for $n[1 - \overline{F}(x;n)] \geq .75$ but only moderately far apart if $n[1 - \overline{F}(x;n)] \leq .5$. They are very close together when $n[1 - \overline{F}(x;n)] \leq .15$.

Similarly, upper and lower bounds are developed for $P(X_1 \leq x)$ in terms of $n\overline{F}(x;n)$, the approximate expression is about halfway between the bounds for cases of principal interest, and these results are usable only if $n\overline{F}(x;n) < 1$. The properties of the bounds are quite similar to those for X_n.

If $\bar{F}(x;n)$ could be suitably estimated for x such that, say, $n[1 - \bar{F}(x;n)] \le .2$, then $P(X_n \le x)$ could be estimated in this range. Likewise, if $\bar{F}(x;n)$ could be estimated for x such that $n\bar{F}(x;n) \le .2$, then $P(X_1 \le x)$ could be estimated in this range. Unfortunately, estimation of $\bar{F}(x;n)$ by the empirical cumulative distribution function of a usable set of observations (average of their cdfs approximately equals $\bar{F}(x;n)$ in the pertinent tail) provides a discontinuous estimate of a step-function nature. Moreover the set size should be very much larger than n if this kind of estimation is to be meaningful.

Estimates of a continuous nature can often be obtained when n is quite large and $\bar{F}(x;n)$ is continuous (or very nearly so) in the pertinent tail. First, consider the case of X_n. Approximation of $\log_e[1 - \bar{F}(x;n)]$ by an expression of the form $- ax + b$, $(a > 0)$, or one of the other two forms occurring in an "asymptote" of X_n for the sample case (for example, see Gumbel, 1958), should often be possible. Use of one of these three forms of approximation reduces the problem of estimating $P(X_n \le x)$ to estimation of two or three parameters. A weighted type of least-squares procedure is developed for estimation of these parameters.

Similar considerations apply to X_1 when n is large and $\bar{F}(x;n)$ is continuous in its lower tail. The three forms of approximation to $\log_e\bar{F}(x;n)$ are those occurring in an "asymptote" of X_1 for the sample case. The parameters in the form of approximation used can be estimated by a least-squares procedure similar to that developed for X_n.

The cdf $\bar{F}(x;n)$ plays an important role for investigations involving order statistics of sets of independent observations (also see Walsh, 1959). Thus, investigation of its properties can be of interest. Several kinds of procedures are given for investigating $\overline{F}(x;n)$ in a specified tail.

Extreme upper percentage points of $\overline{F}(x;n)$, also extreme lower percentage points, can be investigated by use of X_n or X_1, respectively. One-sided confidence intervals and tests are easily developed when $\overline{F}(x;n)$ is continuous (or very nearly so) at the percentile considered. These intervals and tests have rather accurate probability levels if the upper percentage points correspond to values of x such that $\overline{F}(x;n) \geq 1 - .2/n$, and the lower percentage points to x such that $\overline{F}(x;n) \leq .2/n$. Even for percentiles this extreme, the probability levels are only bounded, instead of rather accurately determined, when $\overline{F}(x;n)$ is discontinuous at the percentile considered.

One-sided tolerance intervals for $\overline{F}(x;n)$ can be obtained using X_n, or using X_1. Continuity of $\overline{F}(x;n)$ in the pertinent tail is assumed and the lower bound values are used for the probabilities of the tolerance interval relations.

Finally, tests are easily developed for the two null hypotheses: $\overline{F}(x;n) \equiv F_0(x)$, completely specified, in the upper tail; $\overline{F}(x;n) \equiv F_0(x)$ in the lower tail. Tests for the upper tail use X_n, and tests for the lower tail are based on X_1.

It is to be noted that a distribution exists such that X_n is the largest order statistic for a sample of size n from this distribution. Also, a distribution occurs for which X_1 is the smallest value in a sample of size n. The cdf corresponding to X_n is the geometric mean of the cdfs for the separate observations. The cdf corresponding to X_1 is unity minus the geometric mean of the probabilities of exceeding x for the individual observations. However, their multiplicative definitions result in obscure interpretation of the properties for these distributions, except when they can be approximately expressed in terms of $\overline{F}(x;n)$. Also, distributions that are not easily related arise for different order statistics. An additional advantage of expressing the results in terms of $\overline{F}(x;n)$ is that the distributions for X_n, X_1 and, in fact, all of the order statistics can be expressed (approximately) in terms of n and $\overline{F}(x;n)$ for most situations. Moreover the probabilities in excess of .80 for X_n and less than .20 for X_1 are determined with good accuracy when $\overline{F}(x;n)$ is used.

The following section contains the bounds and approximate expressions for $P(X_n \leq x)$ and $P(X_1 \leq x)$, along with some derivations. Least-squares estimation of parameters for the asymptotic distributions is considered in the next section. Then, there is a section devoted to confidence intervals and tests for upper or lower percentage points of $\overline{F}(x;n)$. The next to last section is concerned with tolerance intervals for $\overline{F}(x;n)$, and the last section contains tests of whether $\overline{F}(x;n)$ equals a completely specified cdf in a stated tail.

BOUNDS AND APPROXIMATE EXPRESSIONS

When $n[1 - \overline{F}(x;n)] < 1$, sharp upper and lower bounds for $P(X_n \leq x)$ are given by

$$1 - n[1 - \overline{F}(x;n)] \leq P(X_n \leq x) \leq \overline{F}(x;n)^n.$$

Sharp upper and lower bounds for $P(X_1 \leq x)$ are provided by

$$1 - [1 - \overline{F}(x;n)]^n \leq P(X_1 \leq x) \leq n\overline{F}(x;n)$$

when $n\overline{F}(x;n) < 1$.

For $n[1 - \overline{F}(x;n)] \leq .25$, the value of $\overline{F}(x;n)^n$ approximately equals

$$1 - n[1 - \overline{F}(x;n)] + (1/2)\{n[1 - \overline{F}(x;n)]\}^2$$

and is exceeded by this value. Thus, in that range for x, this expression can be used, without much loss of sharpness, as the upper bound for $P(X_n \le x)$. The arithmetic average of the lower bound and this upper bound is

$$1 - n[1 - \overline{F}(x;n)] + (1/4)\{n[1 - \overline{F}(x;n)]\}^2 = \{1 - (1/2)n[1 - \overline{F}(x;n)]\}^2,$$

which is the approximate expression for $P(X_n \le x)$. Similarly, for $n\overline{F}(x;n) \le .25$,

$$n\overline{F}(x;n) - (1/2)[n\overline{F}(x;n)]^2$$

is a lower bound that approximately equals the sharp lower bound and

$$n\overline{F}(x;n) - (1/4)[n\overline{F}(x;n)]^2$$

is the approximate expression for $P(X_1 \le x)$.

The sharp upper and lower bounds are derived for $P(X_n \le x)$. A similar derivation would yield sharp upper and lower bounds for $P(X_1 > x) = 1 - P(X_1 \le x)$, and thus for $P(X_1 \le x)$.

Let $F_i(x)$ denote the cdf for the i-th of the n independent observations. Then, $P(X_n \le x)$ equals

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

$$= \exp\{-\sum_{i=1}^{n} \sum_{j=1}^{\infty} [1 - F_i(x)]^j/j\}$$

$$= \exp\{-\sum_{j=1}^{\infty} j^{-1} \sum_{k=0}^{j} \binom{j}{k} [1 - \overline{F}(x;n)]^{j-k} \sum_{i=1}^{n} [\overline{F}(x;n) - F_i(x)]^k\}.$$

For $n[1 - \overline{F}(x;n)] < 1$ and $k \ge 2$,

$$\sum_{i=1}^{n} [\overline{F}(x;n) - F_i(x)]^k$$

is maximum when all but one of the $F_i(x)$ are unity and remaining one is chosen so their arithmetic average is $\bar{F}(x;n)$. This implies that the remaining $F_i(x)$ equals $1 - n[1 - \bar{F}(x;n)]$ and that

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

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

Thus, $P(X_n \leq x)$ is at least equal to (equality possible)

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

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

$$= \exp\left\{\log_e\left(1 - n[1 - \bar{F}(x;n)]\right)\right\} = 1 - n[1 - \bar{F}(x;n)],$$

since

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

which equals n^j.

The sharp upper bound for $P(X_n \leq x)$ is obtained by noticing that, since the geometric mean at most equals the arithmetic mean, the probability is maximized when all the $F_i(x)$ equal $\bar{F}(x;n)$. Thus, this upper bound is $\bar{F}(x;n)^n$.

Alternative proofs for both bounds, also requiring some effort, could be developed from the results of (Hoeffding, 1956).

ESTIMATION OF ASYMPTOTIC PARAMETERS

The other two forms of parametric asymptotic distributions can be transformed into that where $\log_e[1 - \bar{F}(x;n)]$ is approximated by $-ax + b$ for the case of X_n, and where the approximation is of the form $ax + b$ for the case of X_1, with $a > 0$. These transformations are accomplished by suitable changes of variable (for example, see Walsh, 1965). The change of variable involves an additional parameter for the third asymptote but a modification allows a, b, and this parameter to all be estimated by the weighted least-squares procedure. Estimates are explicitly developed only for the case of X_n, but a similar procedure yields estimates for the case of X_1.

Let n' denote the smallest number of observations (with cdfs that average to $\bar{F}(x;n)$ in the upper tail) such that $-ax + b$ is an acceptable approximation to $\log_e[1 - \bar{F}(x;n)]$. Often, the value of n' is unknown but a value that should exceed n' is available. Then, this larger value is used for n'.

Suppose that previous data for use in estimating $\underline{a}$ and $\underline{b}$ consist of several sets of observations (whose cdfs average, approximately, to $\bar{F}(x;n)$ in the upper tail for each set). Pool these observations into one set whose size is denoted by m (ordinarily much larger than the value used for n'). Let $G_m(x)$ be the empirical cdf of these observations while x_L is the smallest x such that $n'[1 - G_m(x)] \leq 1$ and x_M is the smallest x such that $n[1 - G_m(x)] \leq .1$. Finally, for any observation y_c of the total set $y_1, \ldots, y_m$ such that $y_c \leq x_L$, a weight $w(x_M - y_c)$ is specified. Ordinarily, the value of $w(x_M - y_c)$ decreases as the value of $x_L - y_c$ becomes more distant from zero, since the x in the vicinity of x_M are of principal interest.

The estimates of $\underline{a}$ and $\underline{b}$ are the values which minimize

$$(1) \qquad \sum_{y_c \geq x_L} w(x_M - y_c)\{\log_e[1 - G_m(y_c)] + ay_c - b\}^2.$$

They are easily determined by setting the partial derivatives with respect to a and to b equal to zero and simultaneously solving these two equations for a and b.

Additional effort is required when the third asymptote was transformed. Then, the estimates for a and b provide estimates for two parameters of the third asymptote but a third parameter is also involved (occurs explicitly in the change of variable). This parameter is estimated iteratively, with its value being that which minimizes the value of (1) that is first minimized with respect to a and b. That is, starting with an initial guess, values are considered for this third parameter and the corresponding minimum values of (1) with respect to a and b are determined. Ultimately, perhaps using interpolation among the cases already considered, a close approximation to the minimizing value of the third parameter is determined. This approach is also usable when the second asymptote is considered to have three parameters rather than the usual two. Of course, the estimates for a and b are those corresponding to the minimizing value of the third parameter.

INVESTIGATION OF EXTREME PERCENTILES

Let θ_p denote the 100p percent point of $\bar{F}(x;n)$. The principal interest is in the case where $p \geq 1 - .2/n$ and the case where $p \leq .2/n$.

First, consider cases where $\bar{F}(x;n)$ is continuous (or very nearly so) at θ_p. For $p \geq 1 - .2/n$,

$$P(X_n \leq \theta_p) \doteq [1 - n(1 - p)/2]^2.$$

Thus, the interval (X_n, ∞) is a one-sided confidence interval for θ_p with confidence coefficient approximately equal to $[1 - n(1 - p)/2]^2$. For $p \leq .2/n$,

$$P(X_1 \leq \theta_p) \doteq np - (1/4)(np)^2$$

and provides an approximate one-sided confidence interval for θ_p.

Other ranges of p such that $p > 1 - 1/n$ or $p < 1/n$ could be considered in obtaining one-sided confidence intervals. Even for continuity at θ_p, however, the bounds on the confidence coefficient are only moderately close together for $1 - .5/n \leq p < 1 - .2/n$ using X_n, or for $.2/n < p \leq .5/n$ using X_1. They are rather far apart in other cases.

If $\overline{F}(x;n)$ is not continuous at θ_p, the values of $P(X_n \leq \theta_p)$ and $P(X_1 \leq \theta_p)$ are at least equal to the values for continuity at θ_p. The values of $P(X_n < \theta_p)$ and $P(X_1 < \theta_p)$ are at most equal to the values of $P(X_n \leq \theta_p)$ and $P(X_1 \leq \theta_p)$, respectively, for continuity at θ_p.

One-sided tests of the form $\theta_p = \theta_o$, with θ_o a stated number, can be obtained in the usual straightforward manner. Also, X_n is an approximate median estimate of θ_p when $p = 1 - .586/n$, and X_1 is an approximate median estimate of θ_p when $p = .586/n$.

ONE-SIDED TOLERANCE INTERVALS

The one-sided tolerance intervals developed for $\overline{F}(x;n)$ have nonfixed endpoint X_n or X_1. Only cases where $\overline{F}(x;n)$ is continuous (or very nearly so) in the pertinent tail are considered.

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

$$P[\overline{F}(X_n;n) > p] = P(X_n > \theta_p)$$

and hence is at least equal to $1 - p^n$, and at most equal $n(1 - p)$, for $p > 1 - 1/n$. Likewise, the probability that (X_1, ∞) contains at least 100p percent of the probability for $\overline{F}(x;n)$ is

$$P[1 - \overline{F}(X_1;n) > p] = P[\overline{F}(X_1;n) \leq 1 - p]$$

$$= P(X_1 \leq \theta_{1-p}),$$

and this is at least equal to $1 - p^n$, and at most $n(1 - p)$, for $p > 1 - 1/n$.

For given p and a minimum value γ for the probability that the coverage exceeds p, the smallest value for n is determined from $1 - p^n \geq \gamma$, so that $n \geq [\log(1 - \gamma)]/\log p$, which is the value for the case of a random sample of size n from a continuous population.

HYPOTHESIS OF SPECIFIED DISTRIBUTION

Let $\overline{F}(x;n)$ be completely specified as $F_0(x)$ in the upper tail, or the lower tail, by the null hypothesis. Here, the upper tail is for x such that $F_0(x) > 1 - 1/n$, and the lower tail is for x such that $F_0(x) < 1/n$.

Suppose that the principal interest is in disagreement between $\overline{F}(x;n)$ and $F_0(x)$ for the interval of x values in the upper tail such that $p_1 < F_0(x) \leq p_2$. Here, p_1 and p_2 are attainable values for $F_0(x)$ and, under the null hypothesis, $1 - P[p_1 < F_0(X_n) \leq p_2]$ has a small value that is suitable for significance level. Then, the test that rejects $\bar{F}(x;n) = F_o(x)$ in the upper tail if and only if $p_1 < F_o(X_n) \le p_2$ does not hold has significance level equal to the null value of

$$1 - P[F_o(X_n) \le p_2] + P[F_o(X_n) \le p_1],$$

Upper and lower bounds are available for this significance level. These bounds are quite close together if $p_1 \ge 1 - .2/n$. Then, the approximate expression for the null value of $P[F_o(X_n) \le p\ ]$ can be used.

A similar kind of test is easily developed for the lower tail and X_1 used. Also, the critical region could be based on two or more disjoint intervals rather than just one interval. In addition, p_2 could be unity for the upper tail and p_1 could be zero for the lower tail (one-sided cases).

REFERENCES

Gumbel, Emil J., Statistics of Extremes, Columbia University Press, 1958.

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., Handbook of Nonparametric Statistics, II, D. Van Nostrand Co., 1965, Chapter 7.

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

Approximate Distributions for Largest and for Smallest of a Set of Independent Observations

CRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

THOR(S) (First name, middle initial, last name)

John E. Walsh

ORT DATE September 18, 1968	7a. TOTAL NO. OF PAGES 12	7b. NO. OF REFS 4
NTRACT OR GRANT NO. N00014-68-A-0515 OJECT NO. NR 042-260	9a. ORIGINATOR'S REPORT NUMBER(S) 10. 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

ISTRIBUTION STATEMENT

No limitations

UPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

BSTRACT

Consider n independent univariate observations with possibly different distributions. Let X_n and X_1 denote the largest and smallest observations, respectively, while $\overline{F}(x;n)$ is the arithmetic average of the cumulative distribution functions for the individual observations. Approximate expressions, also sharp upper and lower bounds, are developed for $P(X_n \le x)$ and $P(X_1 \le x)$ in terms of n and $\overline{F}(x;n)$. These results are applicable for $P(X_n \le x)$ when x is such that $n[1 - \overline{F}(x;n)] < 1$, and for $P(X_1 \le x)$ when $n\overline{F}(x;n) < 1$. The approximate expressions are reasonably near the bounds if $n[1 - \overline{F}(x;n)] \le .25$ and $n\overline{F}(x;n) \le .25$. Relative error is less than one percent if $n[1 - \overline{F}(x;n)] \le .17$ and $n\overline{F}(x;n) \le .17$; then, $P(X_n \le x) \ge .83$ and $P(X_1 \le x) \le .17$. All possible distributions and all $n \ge 1$ are allowable. Suitable estimation of $\overline{F}(x;n)$ provides estimates of $P(X_n \le x)$ and $P(X_1 \le x)$. In particular, for continuity in pertinent tail of $\overline{F}(x;n)$, asymptotic distributions are often obtained and the problem is simplified to estimating at most three parameters. Approximate confidence regions and tests are obtained for sufficiently extreme upper and lower percentage points of $\overline{F}(x;n)$. Also, for continuity in the pertinent tail, approximate tolerance intervals using X_n or X_1 are developed for $\overline{F}(x;n)$. In addition, tests of $\overline{F}(x;n) \equiv F_0(x)$, completely specified, are developed for x in the pertinent tail.

D FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification