AD 677047

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

SAMPLE SIZES FOR APPROXIMATE INDEPENDENCE OF LARGEST AND SMALLEST ORDER STATISTICS

by

JOHN E. WALSH

Technical Report No. 3
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 OF LARGEST AND SMALLEST ORDER STATISTICS

by

JOHN E. WALSH

Technical Report No. 3
Department of Statistics THEMIS Contract

September 3, 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 OF LARGEST AND SMALLEST ORDER STATISTICS

John E. Walsh

Southern Methodist University*

ABSTRACT

Let X_n and X_1 be the largest and smallest order statistics, respectively, of a random sample of size n. Quite generally, X_n and X_1 are approximately independent for n sufficiently large. Minimum n for attaining at least specified levels of independence are developed. Level of independence is measured by the maximum difference between the true values of $P(X_1 \le x_1, X_n \le x_n)$ and the corresponding values assuming independence of X_n and X_1. The results are for small maximum differences (say, at most .02) and apply to all possible distributions for the population sampled. The value of minimum n is the smallest allowable n for the continuous case but can be too large otherwise. Minimum n is finite for all nonzero differences.

INTRODUCTION AND RESULTS

The largest and smallest order statistics of a random sample tend to statistical independence as the sample size increases. That is, consider a random sample of size n and let X_n and X_1 be the largest and smallest order statistics, respectively. Also consider

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

$$P(X_1 \le x_1, X_n \le x_n) - P(X_1 \le x_1)P(X_n \le x_n), \tag{1}$$

which is nonnegative. As $n \to \infty$, the maximum of this difference (over x_1 and x_n) tends to zero.

Since any n used is finite, there can be interest in how the maximum difference of (1) is affected by n. More specifically, for a given value of the maximum difference, what is the minimum n such that this value is not exceeded? For example, what is the minimum n such that the maximum difference is at most .001? When the maximum difference is small, there is little error in using $P(X_1 \le x_1)P(X_n \le x_n)$ as the joint cumulative distribution function (cdf) for X_n and X_1.

The expression developed for minimum n is based on approximations but is very accurate when the stated maximum difference is small (say, at most .02). This expression provides the smallest permissible value of n when the population sampled is continuous. A smaller value of n could possibly be allowable when the population cdf F(x) is discontinuous, since $F(x_1)$ and/or $F(x_n)$ might not be able to have the values that maximize (1).

Let δ be the specified value for the maximum difference. At most this value occurs if

$$n \ge \frac{-1}{2\log_e(1 - \delta e^2)}\{1 + [1 - 4\log_e(1 - \delta e^2)]^{1/2}\}$$

$$\doteq (1/\delta)e^{-2} + 1/2 \doteq .1353/\delta + .5, \quad (\delta \le .01).$$

For example, the maximum difference is at most .005 if $n \ge 28$.

These results, which are applicable for all possible F(x), again show that X_n and X_1 tend to independence as $n \to \infty$. That is, no matter how small δ is, there are values of n such that the maximum difference is less than δ (say, at most $\delta/2$).

DERIVATIONS

Let $a = a(x_n)$ and $b = b(x_1)$ be defined by $P(X_n \le x_n) = e^{-a}$, $P(X_1 \le x_1) = 1 - e^{-b}$. In the derivations, all values of a and b in the range zero to infinity are considered to be possible (corresponds to the continuous case). Then,

$$F(x_n) = e^{-a/n}, \qquad 1 - F(x_1) = e^{-b/n},$$

so that, in general,

$$P(X_1 \le x_1, X_n \le x_n) = F(x_n)^n - [F(x_n) - F(x_1)]^n$$
$$= e^{-a} - (e^{-a/n} - 1 + e^{-b/n})^n.$$

If X_n and X_1 are independent,

$$P(X_1 \le x_1,\ X_n \le x_n) = F(x_n)^n - F(x_n)^n[1 - F(x_1)]^n$$
$$= e^{-a} - e^{-(a+b)}$$

Thus, the value of (1), the difference of these two probabilities, can be expressed as

$$e^{-(a+b)} - e^{-a}[1 - e^{a/n} + e^{(a-b)/n}]^n$$

which, by some expansions in terms of 1/n, equals

$$e^{-(a+b)} - e^{-a}\exp[-b - ab/n - ab(a+b)/2n^2 + O(1/n^3)]$$
$$\doteq e^{-(a+b)}\{1 - \exp[-ab/n - ab(a+b)/2n^2]\}$$

for n sufficiently large (say, $n \ge 8$) and $a + b$ not large. It is to be noted that $a + b \le -\log_e\delta$ in all cases where the difference is to be at most δ.

This expression is set equal to δ, ($\delta \leq .02$), and the n (not necessarily an integer) yielding this value is determined. Then, this expression for n is maximized with respect to a and b.

First, consider the more crude approximation where terms of order $1/n^2$ are neglected. Then,

$$e^{-(a+b)}(1 - e^{-ab/n}) = \delta$$

so that

$$n \doteq - ab/\log_e(1 - \delta e^{a+b})$$

$$\doteq (1/\delta)abe^{-(a+b)}.$$

Thus, to this order of approximation, a = b = 1 are the maximizing values. That is, the true maximizing values for a and b should be near unity.

Now consider the approximation where terms of order $1/n^3$ are neglected. This yields the quadratic equation

$$n^2 + nab/\log_e(1 - \delta e^{a+b}) + ab(a+b)/2\log_e(1 - \delta e^{a+b}) = 0 ,$$

with solution

$$2n = -[ab/\log_e(1 - \delta e^{a+b})]$$

$$\times\{1 + [1 - 2(a+b)(ab)^{-1}\log_e(1 - \delta e^{a+b})]^{1/2}\}.$$

Expansion with respect to δ yields

$$n\delta = abe^{-(a+b)}[1+(1/2)\delta e^{a+b}]^{-1}[1 + (a+b)(2ab)^{-1}\delta e^{a+b}] + O(\delta^2),$$

so that $\log_e n\delta$ equals

$$\log_e a + \log_e b - a - b - (1/2)\delta e^{a+b} + (a+b)(2ab)^{-1}\delta e^{a+b} + O(\delta^2).$$

This montonically increasing function of n is maximized with respect to a by setting $\partial \log_e n\delta/\partial a$ equal to zero, yielding

$$1/a - 1 - (1/2)\delta e^{a+b}[1 - (a+b)/ab - 1/ab + (a+b)/a^2 b] + O(\delta^2) = 0.$$

Let the terms of order δ^2 be neglected. Also, since δ is small, the solution for the case where terms of order $1/n^2$ are neglected should be usable in the coefficient of δ. This yields the solution $a = 1$, and a similar analysis yields the solution $b = 1$. Thus, $a = b = 1$ is the maximizing choice (to a good approximation) even when terms of order $1/n^2$ are included. Use of $a = b = 1$, combined with n being an integer, yields the expression stated for determining minimum n for given δ.