THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

SAMPLE-LIKE DISTRIBUTION OF AN ORDER STATISTIC UNDER GENERAL NONSAMPLE CONDITIONS AND SOME ASYMPTOTIC IMPLICATIONS

by

John E. Walsh

Technical Report No. 18
Department of Statistics THEMIS Contract

October 7, 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-LIKE DISTRIBUTION OF AN ORDER STATISTIC UNDER GENERAL NONSAMPLE CONDITIONS AND SOME ASYMPTOTIC IMPLICATIONS

John E. Walsh
Southern Methodist University*

ABSTRACT

Consider n univariate observations having an arbitrary joint distribution. In general, the distribution of any order statistic of these observations is shown to be the same as that of this order statistic for a random sample of size n (from a distribution determined by the joint distribution). Thus, individual order statistics can be considered to arise from samples. However, the distribution "sampled" can change greatly with the order statistic. These results are useful in determining asymptotic distributional properties of extremes and percentage points of the observations. That is, for given large n, an asymptotic distribution developed assuming a sample is usable for the more general situation if the distribution "sampled" has a suitable form. Thus, for the continuous case, observed percentage points have asymptotically normal distributions under very general conditions. Also, asymptotic distributions developed for extremes of samples should often be usable for continuous situations. Applications of these asymptotic results for prediction are discussed for situations where several sets of observations (same n for each set) are independently obtained from approximately the same source.

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

INTRODUCTION AND DISCUSSION

The random sample assumption is very convenient in developing distribution properties for order statistics of a set of univariate observations. Thus, in applications, the assumption is often used when it is grossly violated (perhaps involving strong dependence among some observations). Despite this misuse, empirical evidence indicates that at least some of the properties are like those occurring for samples. For instance, a form of asymptotic distribution for the largest sample value seems satisfactory for some practical situations where the sample condition is strongly violated (Gumbel, 1958).

When the observations are independent, an explanation can be given for the applicability of at least some kinds of results developed for samples. Under moderately general circumstances, an order statistic behaves approximately as if it occurred in a random sample (same size) from the distribution equaling the arithmetic average of the distributions for the n observations (Walsh, 1959, 1964). Approximate one-sided and two-sided confidence intervals and tests that have a sign-test nature can be obtained for the percentiles (called generalized percentage points) of this average distribution. Also, for a large set, asymptotic results based on a sample from the average distribution are applicable for extremes.

The establishment and use of distribution properties is severely complicated by the occurrence of nontrivial dependence. However, this paper shows that each order statistic can be treated as if it occurred in a random sample of size n. Unfortunately, due to the various dependencies, the distribution "sampled" can be greatly different for the various order statistics, and there does not seem to be any generally useful relationship among these distributions. In fact, except for the largest and the smallest observation, the distribution "sampled" is not very well identified in terms of individual (conditional or unconditional) distributions for the individual observations.

These sample-like distribution properties have at least one valuable implication. They show that empirical verification of sample-like distributions for order statistics is not a verification of the sample assumption. Distributions of this nature can occur (exactly) for any kind of nonsample situation.

Another worthwhile use is for prediction in a special kind of situation. Here, a new set of univariate observations of size n is to be independently obtained from a source that is approximately the same as several sources which have already provided independent sets of size n. That is, a new multivariate observation with n coordinates is to be independently obtained from a multivariate distribution that is approximately the same as several multivariate distributions which have previously yielded independent multivariate observations with n coordinates. Then, for a given order statistic of the coordinate values, the distribution "sampled" should be nearly the same for all these observations. The prediction problem is to estimate probabilistic properties of a stated order statistic of the new observation from the values of this order statistic for the observations already taken.

The most useful applications of sample-based results seem to occur for asymptotic cases (n large). Then, asymptotic distributions that are completely specified except for values of a very few parameters are eligible for use. The corresponding order statistics from several past multivariate observations provide the basis for estimating these parameters. Probability properties for the new order statistic can be estimated by using the asymptotic distribution with these estimates for the parameters. For example, this approach is useful in conjunction with asymptotic distributions for extremes when the situation is continuous.

A more precise approach is available when a percentage point of the coordinates for a multivariate observation is considered and the situation is continuous. The asymptotic distribution of such a percentage point should ordinarily be approximately normal. Thus, these observed percentage points for the past observations are approximately a random sample from a normal population. The new percentage point is an independent value from approximately the same population. A t-statistic can be used to estimate the probability that the new percentage point will exceed a specified value. Also, whether the new percentage point is from approximately the same population as the percentage points already obtained can be tested. In fact, any procedure for investigating a new observation on the basis of an independent sample from the same normal population is approximately applicable.

The remaining section contains proof that order statistics for a general nonsample situation have sample-like distributions. The distribution "sampled" for a given order statistic is determined by n and the joint distribution of the observations. It can be stated in terms of conditional and unconditional probability distributions for the individual univariate observations. However, the expressions are quite complicated except for the cases of the largest observation and the smallest observation.

VERIFICATION

Let X_t be the t-th order statistic in a set of n observations that can have any possible joint distribution (t = 1 corresponds to the smallest observation, etc.). The cumulative distribution function of X_t for this situation is determined by the joint distribution and is denoted by $F_t(x;n)$. Let $G_t(x;n)$ be determined by

$$\sum_{i=t}^{n} \binom{n}{i} G_t(x;n)^i [1-G_t(x;n)]^{n-i} \equiv F_t(x;n) \quad . \tag{1}$$

The function $G_t(x;n)$ has the properties of a cumulative distribution function (that is, $G_t(x;n)$ is monotonically increasing, $G_t(-\infty;n) = 0$, $G_t(\infty;n) = 1$, etc.)

Consider the t-th order statistic in a random sample of size n from the population with cumulative distribution function $G_t(x;n)$. This distribution is given by the lefthand side of equation (1). Thus, X_t has the same distribution as the t-th order statistic in a random sample of size n from the cumulative distribution function $G_t(x;n)$.

For t = 1 and n, the function $G_t(x;n)$ is easily stated in terms of unconditional and conditional distributions of individual observations.

Also, it can be stated explicitly in terms of $F_t(x;n)$. That is,

$$G_n(x;n) = [F_n(x;n)]^{1/n} ,$$

$$G_1(x;n) = 1 - [F_1(x;n)]^{1/n} .$$

Let $H_j(x;s_{j-1})$ be the conditional probability that the j-th univariate observation has a value at most equal to x given that the first, second, ..., (j-1)th observations have values $\leq x$. Here, s_o indicates no conditions. Then,

$$F_n(x;n) = \prod_{j=1}^{n} H_j(x;s_{j-1})$$, which expresses $G_n(x;n)$ in terms of one unconditional and n-1 conditional cumulative distribution functions. However, the index j could be assigned to the observations in any way. Thus, there are n! ways of expressing $F_n(x;n)$ in terms of one unconditional and n-1 conditional cumulative distribution functions, plus an unlimited number of combinations of these n! expressions.

Finally, let $I_j(x;s'_{j-1})$ be the conditional probability that the j-th observation has a value $\leq x$ given that the first, second, ..., (j-1)th observations have values $>x$. Here, s'_o indicates no conditions. Then,

$$F_1(x;n) = 1 - \prod_{j=1}^{n} [1-I_j(x;s'_{j-1})] ,$$

and $F_1(x;n)$ is expressed in terms of one unconditional and n-1 conditional cumulative distribution functions.

REFERENCES

Gumbel, Emil J., *Statistics of extremes*, Columbia Univ. Press, 1958.

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.