THE DISTRIBUTION OF EXTREME VALUES IN SAMPLES WHOSE MEMBERS ARE SUBJECT TO A MARKOFF CHAIN CONDITION

By Benjamin Epstein

Department of Mathematics, Wayne University

1. Introduction. The extreme value problem as treated in the literature concerns itself with the following question: To find the distribution of the smallest, largest, or more generally the νth largest, or νth smallest values in random samples of size n, drawn from a distribution whose probability law is given by the d.f. $F(x)$. In this formulation the observed sample values $x_1, \cdots, x_n$ are assumed to be statistically independent. While the assumption of independence may be a good approximation to the true state of affairs in some cases, there are situations where this assumption is not justified.

Suppose, for instance, that the observations in the sample are ordered in time. Then it may happen that successive observations are stochastically dependent, the extent of this dependence being a function of the time interval separating these observations.[1] In such cases the present distribution theory for extreme values in samples of size n is inadequate and must be replaced by more general results.

It is clear that a clean-cut analytic solution to the problem of the distribution of extreme values in samples whose members may be stochastically dependent can be expected only for certain special kinds of dependence among successive observations. We are able, in this paper, to obtain the distribution of smallest, largest, second smallest, and second largest values in samples of size n drawn at equally spaced time intervals from a stationary Markoff process.

2. The distribution of smallest and largest values in samples of size n drawn at equally spaced time intervals from a stationary Markoff process. In this section the following assumption is made:

(A) observations $x_1, x_2, \cdots, x_n, \cdots$ are taken in order at times $t = 1, t = 2, \cdots, t = n, \cdots$ from a stationary Markoff random process.

The only information needed in the investigation of a stationary Markoff process at integral values of time is the function

$$F_2(x, y) = \text{Prob } (x_i \leq x, x_{i+1} \leq y), \tag{1}$$

independently of i, where $F_2(x, y)$ must be such that the marginal distribution obtained by integrating over x or y (if x_i or x_{i+1} take on a continuous range of

[1] If the observations $x_1, x_2, \cdots, x_n, \cdots$ are taken at discrete times $t_1, t_2, \cdots, t_n, \cdots$ a measure of stochastic dependence between x_i and x_j is the ordinary coefficient of correlation r_{ij}. If the observations are taken from a continuous stochastic process a natural measure of stochastic dependence between observations made at two different times is the covariance function of the process. In this paper we shall limit ourselves to processes which are discrete in time.

values) or summing over the possible values of x_i or x_{i+1} (if x_i and x_{i+1} can take on only discrete values) is of the form

$$F_1(x) = \text{Prob } (x_i \leq x), \tag{2}$$

independently of i.

An example of a random process meeting condition A is furnished by the Ornstein-Uhlenbeck process [1; 2]. In this case the joint d.f. of x_i and x_{i+1} is given by a non-singular bivariate Gaussian distribution. The results in the present paper are stated completely in terms of the d.f.'s $F_2(x, y)$ and $F_1(x)$ defining the stationary Markoff process and will in particular be valid for observations taken at uniformly spaced time intervals from an Ornstein-Uhlenbeck process.

In this section we shall find the distribution of smallest and largest values in samples $x_1, x_2, \cdots, x_n$ drawn from a random process under assumption A and specified by the bivariate d.f. $F_2(x, y)$ and the associated one dimensional marginal d.f. $F_1(x)$. We first prove Theorem I.

THEOREM I. *Under assumption A, the distribution of largest values in samples of size n is given by the* d.f. $G_n^{(1)}(x) = [F_2(x, x)]^{n-1}/[F_1(x)]^{n-2}$.

To prove this result we note that $G_n^{(1)}(x)$, the probability that the largest value in samples of size n is $\leq x$, is given by

$$G_n^{(1)}(x) = \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_n \leq x). \tag{3}$$

To evaluate the right-hand side of (3) we proceed as follows:

$$\text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_n \leq x) = \tag{4}$$
$$\text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_{n-1} \leq x) \text{ Prob } (x_n \leq x \mid x_1 \leq x, \cdots, x_{n-1} \leq x).$$

But under assumption A, (4) becomes

$$\text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_n \leq x) = \tag{5}$$
$$\text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_{n-1} \leq x) \text{ Prob } (x_n \leq x \mid x_{n-1} \leq x)$$

or

$$G_n^{(1)}(x) = G_{n-1}^{(1)}(x) \text{ Prob } (x_n \leq x \mid x_{n-1} \leq x). \tag{5'}$$

But according to assumption A, and (1) and (2)

$$\text{Prob } (x_n \leq x \mid x_{n-1} \leq x) = \text{Prob } (x_{n-1} \leq x, x_n \leq x)/\text{Prob } (x_{n-1} \leq x) \tag{6}$$
$$= F_2(x, x)/F_1(x).$$

Therefore

$$G_n^{(1)}(x) = G_{n-1}^{(1)}(x)\, F_2(x, x)/F_1(x) \tag{7}$$
$$= G_1^{(1)}(x)\, (F_2(x, x))^{n-1}/(F_1(x))^{n-1}$$
$$= (F_2(x, x))^{n-1}/(F_1(x))^{n-2}.$$

This proves Theorem I.

For $n = 1, 2$, and 3 respectively one gets

$$(8) \quad G_1^{(1)}(x) = F_1(x), \quad G_2^{(1)}(x) = F_2(x, x), \quad G_3^{(1)}(x) = (F_2(x, x))^2/F_1(x).$$

THEOREM II. *Under assumption A, the distribution of smallest values in samples of size n is given by the* d.f.

$$(9) \quad H_n^{(1)}(x) = 1 - \frac{[1 - 2F_1(x) + F_2(x, x)]^{n-1}}{[1 - F_1(x)]^{n-2}}.$$

To prove this result we first note that $H_n^{(1)}(x)$, the probability that the smallest value in samples of size n be $\leq x$ is given by,

$$1 - \text{Prob } (x_1 > x, x_2 > x, \cdots, x_n > x).$$

To evaluate $H_n^{(1)}(x)$ we proceed as follows:

$$(10) \quad \text{Prob } (x_1 > x, x_2 > x, \cdots, x_n > x) =$$

$$\text{Prob } (x_1 > x, x_2 > x, \cdots, x_{n-1} > x) \text{ Prob } (x_n > x \mid x_1 > x, \cdots, x_{n-1} > x).$$

But under assumption A, (10) becomes

$$(11) \quad \text{Prob } (x_1 > x, x_2 > x, \cdots, x_n > x) =$$

$$\text{Prob } (x_1 > x, x_2 > x, \cdots, x_{n-1} > x) \text{ Prob } (x_n > x \mid x_{n-1} > x).$$

But

$$(12) \quad \text{Prob } (x_n > x \mid x_{n-1} > x) = \text{Prob } (x_{n-1} > x, x_n > x)/\text{Prob } (x_{n-1} > x).$$

To evaluate Prob $(x_{n-1} > x, x_n > x)$ we note that

$$(13) \quad \text{Prob } (x_{n-1} > x, x_n > x) + \text{Prob } (x_{n-1} \leq x, x_n > x)$$

$$+ \text{Prob } (x_{n-1} > x, x_n \leq x) + \text{Prob } (x_{n-1} \leq x, x_n \leq x) = 1.$$

Also

$$(14) \quad \text{Prob } (x_{n-1} \leq x, x_n > x) + \text{Prob } (x_{n-1} \leq x, x_n \leq x)$$

$$= \text{Prob } (x_{n-1} \leq x),$$

and

$$(15) \quad \text{Prob } (x_{n-1} > x, x_n \leq x) + \text{Prob } (x_{n-1} \leq x, x_n \leq x)$$

$$= \text{Prob } (x_n \leq x).$$

Recalling that

$$(16) \quad F_2(x, x) = \text{Prob } (x_{n-1} \leq x, x_n \leq x)$$

and

$$(17) \quad F_1(x) = \text{Prob } (x_{n-1} \leq x) = \text{Prob } (x_n \leq x)$$

we get

$$\text{Prob } (x_{n-1} > x, x_n > x) = 1 - 2F_1(x) + F_2(x, x). \tag{18}$$

Therefore (10) becomes

$$\text{Prob } (x_1 > x, x_2 > x, \cdots, x_{n-1} > x, x_n > x) = \tag{19}$$
$$\text{Prob } (x_1 > x, x_2 > x, \cdots, x_{n-1} > x)[1 - 2F_1(x) + F_2(x, x)]/(1 - F_1(x)).$$

Applying the recursion formula (19) successively we obtain

$$\text{Prob } (x_1 > x, x_2 > x, \cdots, x_n > x) = \tag{20}$$
$$\text{Prob } (x_1 > x)[1 - 2F_1(x) + F_2(x, x)]^{n-1}/[1 - F_1(x)]^{n-1}$$
$$= [1 - 2\,F_1(x) + F_2(x, x)]^{n-1}/[1 - F_1(x)]^{n-2}.$$

Therefore $H_n^{(1)}(x)$, the probability that the smallest value in samples of size n is $\leq x$, is given by:

$$H_n^{(1)}(x) = 1 - \frac{[1 - 2F_1(x) + F_2(x, x)]^{n-1}}{[1 - F_1(x)]^{n-2}}. \tag{21}$$

This completes the proof of Theorem II.

In particular for $n = 1, 2$, and 3 respectively the d.f.'s of the smallest value in samples of size n are given by:

$$H_1^{(1)}(x) = F_1(x), \qquad H_2^{(1)}(x) = 2F_1(x) - F_2(x, x),$$
$$H_3^{(1)}(x) = 1 - \frac{[1 - 2F_1(x) + F_2(x, x)]^2}{1 - F_1(x)}. \tag{22}$$

3. Distribution of the second largest and second smallest values in samples of size n drawn at equally spaced time intervals from a stationary Markoff process. Under assumption A of Section II we can state the following theorem.

THEOREM III. *Under assumption A the distribution of second largest values in samples of size n, $n \geq 2$, is given by the* d.f. $G^{(2)}(x)$,

$$G_n^{(2)}(x) = [F_2(x, x)]^{n-1}/[F_1(x)]^{n-2}$$
$$+ 2[F_2(x, x)]^{n-2}\{F_1(x) - F_2(x, x)\}/[F_1(x)]^{n-2}$$
$$+ (n - 2)\,[F_2(x, x)]^{n-3}\,\{F_1(x) - F_2(x, x)\}^2/[F_1(x)]^{n-3}(1 - F_1(x)).$$

To prove this result we first note that $G_n^{(2)}(x)$, the probability that the second largest value is $\leq x$, is given by

$$G_n^{(2)}(x) = \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_n \leq x)$$
$$+ \text{Prob } (x_1 > x, x_2 \leq x, x_3 \leq x, \cdots, x_n \leq x)$$
$$+ \text{Prob } (x_1 \leq x, x_2 > x, x_3 \leq x, x_4 \leq x, \cdots, x_n \leq x) + \cdots \tag{23}$$
$$+ \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_{n-2} \leq x, x_{n-1} > x, x_n \leq x)$$
$$+ \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_{n-1} \leq x, x_n > x).$$

According to Theorem I

$$\text{(24)} \qquad \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_n \leq x) = [F_2(x, x)]^{n-1}/[F_1(x)]^{n-2}.$$

It can readily be shown that

$$\begin{aligned} &\text{Prob } (x_1 > x, x_2 \leq x, x_3 \leq x, \cdots, x_n \leq x) \\ &= \text{Prob } (x_1 \leq x, x_2 \leq x, \cdots, x_{n-1} \leq x, x_n > x) \\ &= [F_2(x, x)]^{n-2} \{F_1(x) - F_2(x, x)\}/[F_1(x)]^{n-2}. \end{aligned} \qquad \text{(25)}$$

It can also be shown that each of the remaining $(n - 2)$ terms on the right-hand side of (23) is equal to

$$\text{(26)} \qquad [F_2(x, x)]^{n-3} \{F_1(x) - F_2(x, x)\}^2/[F_1(x)]^{n-3}(1 - F_1(x)).$$

Combining (23), (24), (25), and (26) we get the desired result in Theorem III, i.e.,

$$\begin{aligned} G_n^{(2)}(x) &= [F_2(x, x)]^{n-1}/[F_1(x)]^{n-2} \\ &+ 2[F_2(x, x)]^{n-2} \{F_1(x) - F_2(x, x)\}/[F_1(x)]^{n-2} \\ &+ (n - 2)[F_2(x, x)]^{n-3} \{F_1(x) - F_2(x, x)\}^2/[F_1(x)]^{n-3}(1 - F_1(x)). \end{aligned} \qquad \text{(27)}$$

In a similar way one can prove Theorem IV.

THEOREM IV. *Under assumption A, the distribution of second smallest values in samples of size n, $n \geq 2$, is given by the* d.f. $H_n^{(2)}(x)$.

$$\begin{aligned} H_n^{(2)}(x) &= 1 - \frac{[1 - 2F_1(x) + F_2(x, x)]^{n-1}}{[1 - F_1(x)]^{n-2}} \\ &- 2\frac{[1 - 2F_1(x) + F_2(x, x)]^{n-2}}{[1 - F_1(x)]^{n-2}} \{F_1(x) - F_2(x, x)\} \\ &- (n - 2)\frac{[1 - 2F_1(x) + F_2(x, x)]^{n-3}}{[1 - F_1(x)]^{n-3}} \frac{\{F_1(x) - F_2(x, x)\}^2}{F_1(x)}. \end{aligned} \qquad \text{(28)}$$

REFERENCES

[1] J. L. DOOB, "The brownian movement and stochastic equations," *Annals of Mathematics*, Vol. 43 (1942), pp. 351.

[2] M. C. WANG AND G. E. UHLENBECK, "On the theory of the brownian motion II," *Reviews of Modern Physics*, Vol. 17 (1945), p. 323.