ON SEVERAL STATISTICS RELATED TO EMPIRICAL DISTRIBUTION FUNCTIONS[1]

By Meyer Dwass

Northwestern University

1. Introduction. Let $X_1, \cdots, X_n$ be n independent random variables, each with the same continuous c.d.f., $F(x)$. Let $F_n(x)$ be the empirical c.d.f. of the X_i's. We consider the following random variables,

$$U_n = \mu\{F(t):F_n(t) - F(t) > 0\},$$

$$D_n = \sup_{-\infty < t < \infty} (F_n(t) - F(t)),$$

$$V_n = \inf_{-\infty < t < \infty} \{F(t):F_n(t) - F(t) = D_n\},$$

where $\{F(t): \quad \}$ denotes the set of values of $F(t)$, for which t satisfies the condition after the colon. These are sets in the interval (0, 1). In the definition of U_n, $\mu\{ \quad \}$ means Lebesgue measure. Obviously, there is no loss of generality in supposing that the X_i's are uniformly distributed over (0, 1) and hence

$$(1) \qquad \begin{cases} U_n = \mu\{t:F_n(t) - t > 0\}, \\ D_n = \sup_{0 \leqq t \leqq 1} (F_n(t) - t), \\ V_n = \inf_{-\infty < t < \infty} \{t:F_n(t) - t = D_r\}. \end{cases}$$

In [5], Kac showed that as $n \to \infty$, U_n has an asymptotic distribution which is uniform over (0, 1). A stronger result was recently obtained by Gnedenko and Mihalevič in [4] in which they showed that *for every n*, U_n is uniformly distributed. Birnbaum and Pyke in a forthcoming paper [2] show that for every n, V_n is also distributed uniformly over (0.1). The methods of [2] and [4] are computational and the purpose of this note is to derive the uniform distribution of U_n and V_n by a short method which employs results of E. S. Andersen and a well-known relationship between the Poisson process and uniformly distributed random variables. In Sec. 3, a generalization of these results is given.

2. Proof of uniform distribution of U_n and V_n. The proof depends on two sets of facts. The first refers to the Poisson process. By this we mean the stochastic process, $X(t)$, with independent and stationary Poisson distributed increments, defined for $t \geqq 0$ and such that $X(0) = 0$. For this process, it is well known that given that $X(1) = n$, a positive integer, then the conditional distribution of the discontinuity (jump) points, $t_1 \leqq t_2 \leqq \cdots \leqq t_n$ of $X(t)$, $0 \leqq t \leqq 1$, is that

Received February 11, 1957; revised June 18, 1957.

[1] Sponsored by Air Force Office of Scientific Research AFOSR-TN-57-784, AD148015 Contract No. AF 49(638)-151.

of the ordered values of n independent, uniform random variables. Another way of saying this, somewhat roughly, is that the conditional distribution of the random function $X(t)$, $0 \leqq t \leqq 1$, given that $X(1) = n$, is that of the empirical c.d.f. of n independent, uniform random variables. For a proof of these facts see p. 400 of [3]. The second set of needed facts is contained in a paper of E. S. Andersen [1], namely:

LEMMA (ANDERSEN). *Let $Y_1, Y_2, \cdots$ be independent and identically distributed random variables. Make the definitions*

$$S_0 = 0 \text{ (a.s.)}, \qquad S_i = \sum_{j=1}^{i} X_j,$$

$$L_r = \textit{smallest } i \textit{ for which } S_i = \max(0, S_1, \cdots, S_r).$$

$$N_r = \textit{number of positive terms in } S_1, \cdots, S_r.$$

Then

$$P(L_r = m \mid S_{r+1} = 0) = P(N_r = m \mid S_{r+1} = 0) = \frac{1}{r+1}, \tag{2}$$

for $m = 0, 1, \cdots, r$ if and only if

$$P(S_i = S_{r+1} = 0) = 0, \qquad (i = 1, 2, \cdots, r). \tag{3}$$

We remark that Andersen's results are much more general, but we state them in a form convenient for our applications.

THEOREM 1. U_n *and* V_n *are each distributed uniformly over* (0, 1).

PROOF. Consider the Poisson process $X(t)$, $0 \leqq t \leqq 1$. Divide the interval (0, 1) into the $r + 1$ parts $(0, 1/(r + 1))$, $(1/(r + 1), 2/(r + 1))$, $\cdots$, $(r/(r + 1), 1)$, where $r + 1$ is greater than n and *is a prime number*. (Whenever we state $r \to \infty$ we will understand that $r + 1$ goes through the primes.) The increments of $X(t)$ in these intervals are independent and identically distributed Poisson random variables. We denote these increments by $W_1, W_2, \cdots, W_{r+1}$, respectively, and define $Y_i = W_i - n/(r + 1)$, $i = 1, \cdots, r + 1$. The Y_i's are independent and identically distributed. We want to show that they satisfy (3) of Andersen's lemma. This is so because $S_i = S_{r+1} = 0$ implies that $(r + 1) \cdot X(i/(r + 1)) = ni$. This cannot hold since by the primeness of $r + 1$, n must be a factor of $X(i/r + 1)$, but since $X(t)$ is non-decreasing this would mean $X(i/(r + 1)) = n$, or $r + 1 = i$, a contradiction; thus (3) holds. Under the condition $X(1) = n$, $X(t)$ is distributed like $F_n(t)$, for $s \leqq t \leqq 1$. Hence we can define U_n, V_n for $X(t)$, $0 \leqq t \leqq 1$. We next observe that when $X(1) = n$, then

$$\left| U_n - \frac{N_r}{r+1} \right| < \frac{A}{r+1}, \qquad \left| V_n - \frac{L_r}{r+1} \right| < \frac{B}{r+1}, \tag{4}$$

where A, B are constants which depend on n but not on r. Thus, under the condition $X(1) = n$, both absolute values in (4) converge in probability to zero as $r \to \infty$. Since $N_r/(r + 1)$ and $L_r/(r + 1)$ are asymptotically uniformly distributed over (0, 1) as $r \to \infty$, this completes the proof.

3. Generalization. A generalization of Theorem 1 can be given which may be of interest. Let

$$X_{11}, \cdots, X_{1n_1}; \cdots; X_{k1}, \cdots, X_{kn_k},$$

be $n = n_1 + \cdots + n_k$ independent random variables each uniformly distributed over (0, 1). Let $F^{(1)}(t), \cdots, F^{(k)}(t)$ be the empirical c.d.f.'s of each of the k sets of variables and define

$$F_\rho(t) = \rho_1 F^{(1)}(t) + \cdots + \rho_k F^{(k)}(t), \qquad 0 \leqq t \leqq 1,$$

where $\rho = (\rho_1, \rho_2, \cdots, \rho_k)$, $\rho_i > 0$, $\rho_1 + \rho_2 + \cdots + \rho_k = 1$. In the special case where $\rho_i = n_i/n$, $i = 1, \cdots, k$, then $F_\rho(t)$ is the empirical c.d.f. of the combined set of n variables. Otherwise $F_\rho(t)$ can only be described as a nondecreasing random step function on (0, 1) such that $F_n(0) = 0$, $F_n(1) = 1$. Nevertheless, random variables U_ρ, D_ρ and V_ρ analogous to U_n, D_n and V_n may be defined for $F_\rho(t)$ exactly as was done in (1) for $F_n(t)$; (replace $F_n(t)$ by $F_\rho(t)$ in (1)). In the following theorem we understand them to be so defined.

THEOREM 2. *U_ρ and V_ρ are each distributed uniformly over* (0, 1).

PROOF. Let $X_1(t), X_2(t), \cdots, X_k(t)$ be k independent Poisson processes and define $X(t) = \rho_1 X_1(t) + \cdots + \rho_k X_k(t)$. Then $X(t)$ is also a process with stationary independent increments. Define now $\rho = (\rho_1, \rho_2, \cdots, \rho_k)$,

$$\begin{cases} \tilde{U}_\rho = \mu\{t: X(t) - X(1)t > 0, 0 \leqq t \leqq 1\}, \\ \tilde{D}_\rho = \sup\limits_{0 \leqq t \leqq 1} (X(t) - X(1)t), \\ \tilde{V}_\rho = \inf\limits_{0 \leqq t \leqq 1} \{t: X(t) - X(1)t = \tilde{D}_\rho\}. \end{cases}$$

We suppose first that

$$\rho_1 = a_1/a, \cdots, \rho_k = a_k/a, \tag{5}$$

where $a_1, \cdots, a_k$ are positive integers, and $a_1 + \cdots + a_k = a$. If b is a number such that $P(X(1) = b) > 0$, then $\tilde{U}_\rho$, $\tilde{V}_\rho$ are each uniformly distributed over (0, 1) given that $X(1) = b$. The proof of this fact follows exactly the proof of theorem 1. In particular the definition of the ρ_i's by (5) allows a verification of the condition (3) of Andersen's lemma which is exactly analogous to that done in the proof of Theorem 1. Since the ρ_i's as defined by (5) are dense in the set of all possible ρ_i's, it follows by a simple continuity argument that the conditional distribution of $\tilde{U}_\rho$, $\tilde{V}_\rho$ given that $X(1) = b$, is uniform *without* the restriction (5). If $X(1) = \rho_1 X_1(1) + \cdots + \rho_k X_k(1) = b$, this need not uniquely determine the values of the $X_i(1)$. That is, there may be two different sets of positive or zero integers, $x_1, \cdots, x_k$; $y_1, \cdots, y_k$, such that

$$\rho_1 x_1 + \cdots + \rho_k x_k = \rho_1 y_1 + \cdots + \rho_k y_k = b.$$

On the other hand, there is a dense subset of the k-dimensional unit cube where this cannot happen, namely any dense subset, each point of which has rationally independent coordinates. Thus, in such a dense subset $X(1) = \rho_1 n_1 + \cdots + \rho_k n_k$ if and only if $x_i(1) = n_1, \cdots, x_k(1) = n_k$, for a set of ρ_i's which are dense in the set of all possible ρ_i's. For such ρ_i's the conditional distribution of $\tilde{U}_\rho$ and $\tilde{V}_\rho$ given that $X_1(1) = n_1, \cdots, X_k(1) = n_k$, is thus uniform. This holds also for the exceptional ρ_i's by a continuity argument. This completes the proof since $F^{(1)}(t), \cdots, F^{(k)}(t)$ are distributed like $X_1(t), \cdots, X_k(t)$ for $0 \leqq t \leqq 1$, under the conditions that $X_1(1) = n_1, \cdots, X_k(t) = n_k$.

4. Concluding remarks. The linear combinations of Theorem 2 are convex $(\rho_1 + \cdots + \rho_k = 1)$ and positive $(\rho_i > 0)$. The convexity, as well as the strict positivity, is a matter of convenience. The condition of non-negativeness, however, cannot be removed. It is easy to verify directly, for example, that the theorem does not hold for

$$F_\rho(t) = \rho_1 F^{(1)}(t) + \rho_2 F^{(2)}(t),$$

if $\rho_1 > 0$ and $\rho_2 < 0$. The trouble arises because the condition (3) of Andersen's lemma fails to hold.

REFERENCES

[1] E. S. Andersen, "On the fluctuations of sums of random variables," *Math. Scand.*, Vol. 1 (1953), pp. 263–285.

[2] Z. W. Birnbaum and R. Pyke, "On some distributions related to the statistic D_n^+," Technical report no. 23, July 17, 1956, Laboratory of Statistical Research, Dept. of Math., University of Washington, Seattle, Washington.

[3] J. L. Doob, *Stochastic Processes*, John Wiley and Sons, New York, (1953).

[4] B. V. Gnedenko and V. S. Mihalevič, "Two theorems on the behavior of empirical distribution functions," *Doklady Akad. Nauk SSSR* (N.S.), Vol. 85 (1952), pp. 25–27.

[5] M. Kac, "On deviations between theoretical and empirical distributions," *Proc. Nat. Acad. Sci., U.S.A.*, Vol. 35 (1949), pp. 252–257.