The Annals of Probability
1978, Vol. 6, No. 1, 133-137

A PROOF OF KAKUTANI'S CONJECTURE ON RANDOM SUBDIVISION OF LONGEST INTERVALS

BY W. R. VAN ZWET

University of Leiden

Choose a point at random, i.e., according to the uniform distribution, in the interval (0, 1). Next, choose a second point at random in the largest of the two subintervals into which (0, 1) is divided by the first point. Continue in this way, at the nth step choosing a point at random in the largest of the n subintervals into which the first $(n-1)$ points subdivide (0, 1). Let F_n be the empirical distribution function of the first n points chosen. Kakutani conjectured that with probability 1, F_n converges uniformly to the uniform distribution function on (0, 1) as n tends to infinity. It is shown in this note that this conjecture is correct.

1. Introduction. Let X_1 be uniformly distributed on (0, 1). For $n = 2, 3, \cdots$, the conditional distribution of X_n given $X_1, \cdots, X_{n-1}$ is uniform on the largest of the n subintervals into which $X_1, \cdots, X_{n-1}$ subdivide (0, 1). Let F_n denote the empirical distribution function (df) of $X_1, \cdots, X_n$, thus $F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \leqq x\}}$.

THEOREM. *With probability* 1

$$\lim_{n\to\infty} \sup_{x \in (0,1)} |F_n(x) - x| = 0 . \tag{1.1}$$

At first sight the truth of this statement seems intuitively obvious. The Glivenko–Cantelli theorem tells us that (1.1) holds with probability 1 if X_1, $X_2, \cdots$ are independent and identically distributed (i.i.d.) according to the uniform distribution on (0, 1). Compared with this case, one feels that F_n should converge to the uniform df even faster in the present situation, because at each step one is putting a point where it is needed most, i.e., in the largest subinterval. At the same time, however, it is clear that the procedure by which the points are chosen makes their joint distribution extremely complicated. To be convinced of this, one only has to try and write down what happens in just the first few steps.

The main idea of the proof is the introduction of a stopping rule for which the stopped sequence has an essentially simpler character than the original one. For $t \in (0, 1)$, let N_t be the smallest natural number n for which $X_1, \cdots, X_n$ subdivide (0, 1) into $(n+1)$ subintervals of length $\leqq t$. Correspondingly, define $N_t = 0$ for $t \geqq 1$. The basic property of the stopped sequence $X_1, \cdots, X_{N_t}$ is that any (sub-) interval appearing during its construction will receive another random point before the sequence is stopped, if and only if its length exceeds t. It follows that the joint distribution of N_t and the set $\{X_1, \cdots, X_{N_t}\}$ remains unchanged if at each step the next point is chosen at random in any one of the

Received February 28, 1977.
AMS 1970 *subject classifications*. Primary 60F15; Secondary 60K99.
Key words and phrases. Glivenko–Cantelli type theorem.

existing subintervals of length $> t$ rather than in the largest subinterval as prescribed by the original procedure. In the first place this implies that for $t \in (0, 1)$, the conditional distribution of $(N_t - 1)$ given $X_1 = x$ is that of the sum of the numbers of random points needed to subdivide the intervals $(0, x)$ and $(x, 1)$ independently and in the prescribed way into subintervals of length $\leqq t$. By blowing up these intervals to length 1 and replacing t by t/x and $t/(1 - x)$ respectively one sees that

$$\mathscr{L}(N_t \mid X_1 = x) = \mathscr{L}(N_{t/x} + N_{t/(1-x)} + 1), \qquad 0 < t < 1, \tag{1.2}$$

where for $N_{t/x}$ and $N_{t/(1-x)}$ independent copies are chosen.

Another consequence of the abovementioned property of the stopped sequence is the following. Take $x \in (0, 1)$ and let $N_t(x)$ denote the number of values in $(0, x]$ among $X_1, \cdots, X_{N_t}$, thus $N_t(x) = N_t F_{N_t}(x)$. Suppose that $0 < t < x$ and let ξ be the first value in the interval $[x - t, x]$ occurring in the sequence $X_1, \cdots, X_{N_t}$. If from the $N_t(x)$ values in $(0, x]$ we delete all values in $(\xi, x]$, the number remaining is distributed as the number of random points needed to subdivide $(0, x]$ into subintervals of length $\leqq t$ in the prescribed way, i.e., as $N_{t/x}$. If also $t < 1 - x$, the same argument applied to the interval $(x, 1)$ shows that there exist copies of $N_{t/x}$ and $N_{t/(1-x)}$ such that

$$N_{t/x} \leqq N_t(x) \leqq N_t - N_{t/(1-x)} \tag{1.3}$$

with probability 1. This clearly holds for all t since $N_{t/x} = 0$ for $t \geqq x$ and $N_{t/(1-x)} = 0$ for $t \geqq 1 - x$.

2. Proof of the theorem. For $t \in [\frac{1}{2}, 1)$, the stopped sequence $X_1, \cdots, X_{N_t}$ never returns to a subinterval it has left. Hence the Markov inequality yields

$$P(N_t > k) = P(\textstyle\prod_{i=1}^{k} \{U_i \vee (1 - U_i)\} > t) \leqq (\frac{3}{4})^k t^{-1}, \qquad \frac{1}{2} \leqq t < 1, \tag{2.1}$$

where $U_1, U_2, \cdots$ are i.i.d. with a uniform distribution on $(0, 1)$, so that $E\{U_i \vee (1 - U_i)\} = \frac{3}{4}$. It follows that $EN_t^m < \infty$ for every $m \geqq 0$ and $\frac{1}{2} \leqq t < 1$. For $s, t \in (0, 1)$, N_{st} is stochastically smaller than a sum of $(N_t + 1)$ copies of N_s and hence $EN_t^m < \infty$ for $m \geqq 0$ and $0 < t < 1$. Since EN_t^m is nonincreasing in t,

$$\sup_{t_0 \leqq t < 1} EN_t^m < \infty \qquad \text{for} \quad 0 < t_0 < 1 \quad \text{and} \quad m \geqq 0. \tag{2.2}$$

Clearly $EN_t^m = 0$ for $t \geqq 1$ and $m \geqq 0$ because $N_t = 0$ for $t \geqq 1$. Another consequence of (2.1) is that for $\frac{1}{2} < t < 1$

$$P(N_t > k) \leqq \textstyle\prod_{i=1}^{k} P(\{U_i \vee (1 - U_i)\} > t) \leqq \{2(1 - t)\}^k,$$

$$EN_t = \textstyle\sum_{k=0}^{\infty} P(N_t > k) \leqq \dfrac{1}{2t - 1}.$$

Since $N_t \geqq 1$ a.s. for $t < 1$, it follows that

$$\lim_{t \uparrow 1} EN_t = 1. \tag{2.3}$$

Define $\mu(t) = EN_t$. For $t \geqq 1$, $\mu(t) = 0$ and in view of (1.2) one finds that

for $0 < t < 1$,

$$(2.4)\qquad \mu(t) = \int_0^1 \left\{\mu\left(\frac{t}{x}\right) + \mu\left(\frac{t}{1-x}\right) + 1\right\} dx = 2\int_0^1 \mu\left(\frac{t}{x}\right) dx + 1$$

$$= 2\int_t^1 \mu\left(\frac{t}{x}\right) dx + 1 = 2t\int_t^1 \frac{\mu(y)}{y^2}\, dy + 1 .$$

Now $\sup_{y \geq t} \mu(y) < \infty$ for $t > 0$ because of (2.2) and hence (2.4) implies that μ is continuous and even differentiable on (0, 1) with

$$\left(\frac{\mu(t) - 1}{t}\right)' = \frac{\mu'(t)}{t} - \frac{\mu(t) - 1}{t^2} = -2\frac{\mu(t)}{t^2},$$

or

$$\frac{\mu'(t)}{\mu(t) + 1} = -\frac{1}{t} .$$

Together with (2.3) this yields

$$(2.5)\qquad \mu(t) = \frac{2}{t} - 1 \qquad \text{for} \quad 0 < t < 1 .$$

Let $v(t)$ denote the variance of N_t and apply (1.2) again, this time also using the independence of $N_{t/x}$ and $N_{t/(1-x)}$ in (1.2). In view of (2.5) one obtains for $0 < t \leqq \frac{1}{2}$,

$$v(t) = E\left(N_t - \frac{2}{t} + 1\right)^2$$

$$= \int_0^1 E\left(N_{t/x} + N_{t/(1-x)} - \frac{2x}{t} - \frac{2(1-x)}{t} + 2\right)^2 dx$$

$$= \int_0^1 E\left(N_{t/x} - \frac{2x}{t} + 1\right)^2 dx + \int_0^1 E\left(N_{t/(1-x)} - \frac{2(1-x)}{t} + 1\right)^2 dx$$

$$+ 2\int_0^1 E\left(N_{t/x} - \frac{2x}{t} + 1\right) E\left(N_{t/(1-x)} - \frac{2(1-x)}{t} + 1\right) dx$$

$$= 2\int_0^1 E\left(N_{t/x} - \frac{2x}{t} + 1\right)^2 dx ,$$

where the cross-product term vanishes because of (2.5) and because either $t/x < 1$ or $t/(1-x) < 1$ for $t \in (0, \frac{1}{2}]$ and $x \in (0, 1)$, $x \neq \frac{1}{2}$. So for $t \in (0, \frac{1}{2}]$,

$$(2.6)\qquad v(t) = 2\int_t^1 v(t/x)\, dx + 2\int_0^t \left(\frac{2x}{t} - 1\right)^2 dx = 2t\int_t^1 \frac{v(y)}{y^2} dy + \frac{2t}{3} .$$

Because of (2.2), $\sup_{y \geq t} v(y) < \infty$ for $t > 0$, and together with (2.6) this ensures that v is continuous on $(0, \frac{1}{2}]$ and differentiable on $(0, \frac{1}{2})$ with

$$\left(\frac{v(t)}{t}\right)' = \frac{v'(t)}{t} - \frac{v(t)}{t^2} = -2\frac{v(t)}{t^2}$$

or

$$\frac{v'(t)}{v(t)} = -\frac{1}{t} .$$

Hence, if $c = \frac{1}{2}v(\frac{1}{2})$,

$$v(t) = \frac{c}{t} \qquad \text{for} \quad 0 < t < \tfrac{1}{2}\,. \tag{2.7}$$

For $m = 2, 3, \cdots$, define $M_m = N_{m-2}$ and $M_m(x) = N_{m-2}(x)$ for $x \in (0, 1)$. Then (2.5), (2.7) and the Bienaymé–Chebyshev inequality imply that

$$P(|M_m - 2m^2 + 1| \geqq m^{\frac{5}{3}}) \leqq \frac{\sigma^2(M_m)}{m^{\frac{10}{3}}} = cm^{-\frac{4}{3}}\,.$$

By the Borel–Cantelli lemma

$$\limsup_m m^{-\frac{5}{3}}|M_m - 2m^2| \leqq 1 \quad \text{a.s.}$$

so that

$$\lim_{m\to\infty} \frac{M_m}{2m^2} = 1 \quad \text{a.s.}\,, \tag{2.8}$$

$$\lim_{m\to\infty} \frac{M_{m+1}}{M_m} = 1 \quad \text{a.s.} \tag{2.9}$$

For fixed $x \in (0, 1)$ and $t = m^{-2}$, the reasoning leading to (2.8) may also be applied to each of the three terms on the left- and right-hand sides of (1.3). Since the argument does not involve joint distributions for different values of m, it follows without further specification of the copies chosen in (1.3) that for any fixed $x \in (0, 1)$

$$\lim_{m\to\infty} \frac{M_m(x)}{2m^2} = x \quad \text{a.s.}\,,$$

or, in view of (2.8),

$$\lim_{m\to\infty} F_{M_m}(x) = x \quad \text{a.s.} \tag{2.10}$$

For $M_m \leqq n \leqq M_{m+1}$,

$$\begin{aligned} |F_n(x) - x| &\leqq \frac{M_m}{n}\,|F_{M_m}(x) - x| + \frac{n - M_m}{n}\,\{x \vee (1 - x)\} \\ &\leqq |F_{M_m}(x) - x| + \left(1 - \frac{M_m}{M_{m+1}}\right) \end{aligned}$$

and together with (2.8), (2.9) and (2.10) this implies that for every fixed $x \in (0, 1)$,

$$\lim_{n\to\infty} F_n(x) = x \quad \text{a.s.} \tag{2.11}$$

By a standard argument this yields (1.1) and the theorem is proved.

Acknowledgment. The author recalls with pleasure the 1976 stochastics meeting at Oberwolfach where R. M. Dudley introduced the participants to Kakutani's conjecture and proceeded to shoot down our combined attempts at solving the problem.

Note added in proof. After this paper was submitted it has come to the author's attention that J. Komlós and G. Tusnády had also arrived at the conclusion that Kakutani's conjecture can be proved by the method employed in this paper. More recently essentially the same proof was given again independently in Lootgieter (1977a); an outline of this paper is given in Lootgieter (1977b). For the solution of a related nonrandom problem the reader is referred to Kakutani (1975), Adler and Flatto (1977) and Lootgieter (loc. cit.).

REFERENCES

ADLER, R. L. and FLATTO, L. (1977). Uniform distribution of Kakutani's interval splitting procedure. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **38** 253–259.

KAKUTANI, S. (1975). A problem in equidistribution. *Lecture Notes in Math.* **541** 369–376. Springer, Berlin.

LOOTGIETER, J. C. (1977a). Sur la répartition des suites de Kakutani. Technical Report, Université Paris VI; submitted to *Ann. Henri Poincaré*.

LOOTGIETER, J. C. (1977b). Sur la répartition des suites de Kakutani. *C.R. Acad. Sci. Paris* **285A** 403–406.

INSTITUTE OF APPLIED MATHEMATICS
AND COMPUTER SCIENCE
UNIVERSITY OF LEIDEN
WASSENAARSEWEG 80, LEIDEN
THE NETHERLANDS