Statistics & Probability Letters 10 (1990) 321–324
North-Holland September 1990

ON AN INTERVAL SPLITTING PROBLEM

F. Thomas BRUSS *, S. Rao JAMMALAMADAKA and Xian ZHOU
Statistics and Applied Probability Program, University of California, Santa Barbara, CA 93106, USA

Received October 1988
Revised July 1989

Abstract: Let $X_1, X_2, \ldots,$ be i.i.d. random variables, which are uniformly distributed on [0,1]. Further let $I_1(0) = [0, 1]$ and let $I_k(n)$ denote the kth largest interval generated by the points 0, $X_1, X_2, \ldots, X_{n-1}$, 1 (or equivalently, the interval corresponding to the kth largest spacing at the nth stage). This note studies the question for which classes of sequences $k = k(n)$, will the interval $I_{k(n)}(n)$ be hit (a.s.) only finitely often, as well as infinitely often.

Keywords: Interval splitting, spacing, extended Borel–Cantelli lemma.

1. Introduction

There are several interesting problems concerning the division of an interval into subintervals. We just mention two in connection with what Pyke (1980) calls the uniform model (U-model) and the Kakutani model (K-model).

In the U-model, points are thrown into (or chosen on) [0,1] independently of each other and all according to the uniform distribution function U on [0,1]. Since U is continuous, n points $X_1, X_2, \ldots, X_n$ will split [0,1] a.s. into $n+1$ intervals. The Glivenko–Cantelli result affirms that the empirical distribution function F_n of $X_1, \ldots, X_n$ will tend a.s. to the uniform distribution function U on [0,1].

In what follows, we shall study the question whether in the U-model, certain intervals are hit infinitely often or not. Thus in the following, the random variables $X_1, X_2, \ldots$ are all independent and uniformly distributed over [0,1]. We mentioned, in passing, that the K-Model of Kakutani corresponds to choosing X_n uniformly from the largest interval generated by the preceding observations. See Pyke (1980) and references contained therein.

2. Results

We first show that the sequence of the largest intervals will be hit infinitely often. Let $I_1(n)$ denote the largest of the n intervals generated by $X_1, X_2, \ldots, X_{n-1}$. Further, let i.o. be shorthand for 'infinitely often'. Then we have:

Proposition 1. $P(X_n \in I_1(n)\ i.o.) = 1$.

Proof. This is easy to see: Notice that the largest of these n intervals, which sum up to one, has always length greater than $1/n$. Therefore

$$\sum_n P(X_n \in I_1(n) \mid X_1, X_2, \ldots, X_{n-1}) = \infty$$

and the rest follows from the extended Borel–Cantelli lemma (see e.g. Breiman, 1968, Corollary 5.29). □

It is interesting to see that the smallest interval, on the other hand, will almost surely be hit a *finite* numbers of times only, i.e. if $I_n(n)$ denotes

* On leave from Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium. Now at Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90024, USA.

0167-7152/90/$3.50 © 1990 – Elsevier Science Publishers B.V. (North-Holland)

the smallest of the n intervals generated by $X_1, \ldots, X_{n-1}$, then:

Proposition 2. *$P(X_n \in I_n(n)$ i.o.$) = 0$.*

Proof. Let $\mathrm{Min}(n) = |I_n(n)|$ denote the length of $I_n(n)$ and let

$$p_n(\alpha) = P(\mathrm{Min}(n) > \alpha), \quad 0 < \alpha < 1.$$

Notice that $p_n(\alpha) = 0$ for all $\alpha \in [1/n, 1]$. Further it is easy to see that

$$p_n(\alpha) = p_{n-1}(\alpha) \cdot P(A_n \mid \mathrm{Min}(n-1) > \alpha) \tag{1}$$

where A_n is the event that the nth point X_n does not fall into the α-neighborhood of any of the points $0, X_1, \ldots, X_{n-1}, 1$. Thus

$$P(A_n \mid \mathrm{Min}(n-1) > \alpha) < 1 - n\alpha, \quad 0 < \alpha < 1/n.$$

Using (1), we obtain by recurrence for all $0 < \alpha < 1/n$,

$$p_n(\alpha) < \prod_{k=1}^{n} (1 - k\alpha) \leqslant (1 - \alpha)^{n(n+1)/2} \tag{2}$$

where the right-hand side inequality follows easily by induction using the Bernoulli inequality. Now let $\alpha_n = n^{-\beta}$ where $1 < \beta < 2$. Then we obtain from (2),

$$p_n(\alpha_n) \leqslant (1 - n^{-\beta})^{n^2/2} = (1 - n^{-\beta})^{n^{\beta} n^{2-\beta}/2} \sim (1/\mathrm{e})^{n^{2-\beta}/2}$$

so that with $0 < 2 - \beta < 1$, $\Sigma_n p_n(\alpha_n) < \infty$. The general part of the standard Borel–Cantelli lemma implies therefore $P(\mathrm{Min}(n) > \alpha_n$ i.o.$) = 0$. Thus, for all n sufficiently large, $P(X_n \in I_n(n)$ i.o.$) = 0$ since $\Sigma_n n^{-\beta} < \infty$. This completes the proof. □

We now ask: "How large should this interval be for it to be hit infinitely often?" For instance, would an interval corresponding to the 'median spacing' be hit infinitely often? To get some insight into this, we introduce more formal notations and establish two general results.

Let $X_1, X_2, \ldots,$ be i.i.d. random variables with uniform distribution on [0,1] and $X_{1:n} \leqslant \cdots \leqslant X_{n-1:n}$ be the order statistics of $X_1, \ldots, X_{n-1}$, where $n > 1$. Define the uniform spacings by

$$D_{in} = X_{i+1:n} - X_{i:n}, \quad i = 0, 1, \ldots, n-1,$$

where $X_{0:n} = 0$ and $X_{n:n} = 1$. Let $D_{1:n} \geqslant \cdots \geqslant D_{n:n}$ be the ordered uniform spacings and $I_k(n)$ denote the kth largest interval with length $D_{k:n}$. We are concerned with the probability that the next observation X_n falls into $I_k(n)$, where $k = k(n)$ satisfies $1 \leqslant k(n) \leqslant n$. We need the following result for which we provide a direct proof. See also Theorem 2.2 of Holst (1980).

Lemma 1. *For $1 \leqslant k \leqslant n$,*

$$P(X_n \in I_k(n)) = \binom{n-1}{k-1} \int_0^1 t^{k-1}(1-t)^{n-k}(-\ln t)\, \mathrm{d}t.$$

Proof. Since X_n is uniform on [0,1], clearly

$$P(X_n \in I_k(n)) = E[P(X_n \in I_k(n) \mid X_1, \ldots, X_{n-1})] = E[D_{k:n}] = \int_0^1 x g(x)\, \mathrm{d}x \tag{3}$$

where $g(x)$ is the density of $D_{k:n}$, which is given by

$$g(x) = (n-1)k\binom{n}{k} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} \times \langle 1 - (j+k)x \rangle^{n-2}, \tag{4}$$

with $\langle x \rangle = \max(x, 0)$. For a derivation of (4), see Rao and Sobel (1980), equation (3.6). Since

$$\int_0^1 x\langle 1 - (j+k)x \rangle^{n-2}\, \mathrm{d}x = \int_0^{1/(j+k)} x(1 - (j+k)x)^{n-2}\, \mathrm{d}x = \frac{1}{n(n-1)} \frac{1}{(j+k)^2},$$

from (3) and (4) it follows that

$$P(X_n \in I_k(n)) = \binom{n-1}{k-1} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} \frac{1}{(j+k)^2}. \tag{5}$$

We now rewrite the sum on the right-hand side of (5) in a different form. Since

$$(1-t)^{n-k} = \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} t^j,$$

we have

$$\int_0^s t^{k-1}(1-t)^{n-k}\,dt$$
$$= \sum_{j=0}^{n-k}(-1)^j\binom{n-k}{j}\frac{1}{j+k}s^{j+k}.$$

Hence for $0<\theta<1$,

$$\int_0^\theta \frac{1}{s}\int_0^s t^{k-1}(1-t)^{n-k}\,dt\,ds$$
$$= \sum_{j=0}^{n-k}(-1)^j\binom{n-k}{j}\frac{1}{j+k}\int_0^\theta s^{j+k-1}\,ds$$
$$= \sum_{j=0}^{n-k}(-1)^j\binom{n-k}{j}\frac{1}{(j+k)^2}\theta^{j+k}.$$

Let $\theta\to 1-$. Then by (5) and the above equalities we get

$$P(X_n\in I_k(n))$$
$$=\binom{n-1}{k-1}\int_0^1\frac{1}{s}\int_0^s t^{k-1}(1-t)^{n-k}\,dt\,ds$$
$$=\binom{n-1}{k-1}\int_0^1 t^{k-1}(1-t)^{n-k}\int_t^1\frac{1}{s}\,ds\,dt,$$

which gives the required result because $\int_t^1(1/s)\,ds = -\ln t$. □

Thus it follows from Lemma 1 and the Borel–Cantelli lemma that $P(X_n\in I_k(n)\text{ i.o.})=1$ implies

$$\sum_{n=2}^{\infty}\binom{n-1}{k-1}\int_0^1 t^{k-1}(1-t)^{n-k}(-\ln t)\,dt=\infty \tag{6}$$

and $P(X_n\in I_k(n)\text{ i.o.})=0$ if the series in (6) converges. The following two propositions cover most cases of $k=k(n)$, for which (6) holds or fails, and generalize partly Propositions 1 and 2.

Proposition 3. *If* $\limsup_{n\to\infty}(k(n)/n)<1$, *then* $\sum_n P(X_n\in I_k(n))=\infty$.

Proof. By Taylor expansion of $-\ln t$ around 1 we obtain

$$-\ln t=(1-t)+\tfrac{1}{2}(1-t)^2+\cdots\geqslant 1-t$$

for all $t\in[0,1]$. It follows that

$$\binom{n-1}{k-1}\int_0^1 t^{k-1}(1-t)^{n-k}(-\ln t)\,dt$$
$$\geqslant\binom{n-1}{k-1}\int_0^1 t^{k-1}(1-t)^{n-k+1}\,dt$$
$$=\binom{n-1}{k-1}\frac{(k-1)!(n-k+1)!}{(n+1)!}=\frac{n-k+1}{(n+1)n}.$$

If $\limsup_{n\to\infty}(k(n)/n)<1$, then there are $N>1$ and $c<1$ such that $k\leqslant cn$ for all $n\geqslant N$. Consequently,

$$\sum_{n=2}^{\infty}\frac{n-k+1}{(n+1)n}\geqslant\sum_{n=2}^{\infty}\frac{(1-c)n+1}{(n+1)n}$$
$$>\sum_{n=2}^{\infty}\frac{1-c}{n+1}=\infty.$$

Thus (7) implies that (6) holds, and so the proposition follows. □

Proposition 4. *If* $k\geqslant n-An^p$ *for some* $A\geqslant 0$ *and* $0\leqslant p<1$, *then* $P(X_n\in I_k(n)$ *i.o.*$)=0$.

Proof. It is easy to verify that

$$t(-\ln t)\leqslant 1-t\quad\text{for all } t\in(0,1].$$

(For example, the two sides are equal at $t=1$ and the derivative of $t(-\ln t)$ is greater than that of $1-t$ on (0, 1).) Therefore for $k\geqslant 2$,

$$\sum_{n=2}^{\infty}\binom{n-1}{k-1}\int_0^1 t^{k-1}(1-t)^{n-k}(-\ln t)\,dt$$
$$\leqslant\sum_{n=2}^{\infty}\binom{n-1}{k-1}\int_0^1 t^{k-2}(1-t)^{n-k+1}\,dt$$
$$=\sum_{n=2}^{\infty}\binom{n-1}{k-1}\frac{(k-2)!(n-k+1)!}{n!}$$
$$=\sum_{n=2}^{\infty}\frac{n-k+1}{(k-1)n}.$$

By the condition $k\geqslant n-An^p$, there are $M>0$ and $N\geqslant 2$ such that

$$\frac{n-k+1}{(k-1)n}\leqslant\frac{An^p+1}{(n-An^p-1)n}\leqslant Mn^{-(2-p)}$$

for all $n\geqslant N$.

Thus the last series in (8) converges since $p < 1$, and so does the first series in (8). Hence $P(X_n \in I_k(n) \text{ i.o.}) = 0$ follows from Lemma 1 and the Borel–Cantelli lemma. □

Remark. $k = n$ is a particular case which satisfies the condition of Proposition 4 with $A = 0$. Hence Proposition 4 is a generalization of Proposition 2.

Acknowledgement

We are grateful to Prof. Sidney Port (UCLA) for his comments on an earlier version of this paper. The first author would also like to thank the Alexander von Humboldt Foundation for their financial support.

References

Breiman, L. (1968), *Probability* (Addison-Wesley, Reading, MA).

Holst, L. (1980), On the lengths of the pieces of a stick brocken at random, *J. Appl. Probab.* **17**, 623–634.

Pyke, R. (1980), The asymptotic behaviour of spacing under Kakutani's model for interval subdivision, *Ann. Probab.* **8**, 157–163.

Rao, J.S. and M. Sobel (1980), Incomplete Dirichlet integrals with applications to ordered uniform spacings. *J. Multivariate Anal.* **10**, 603–610.