Some remarks on random numbers

K. R. Parthasarathy

It is shown that the number of k-peaks in a sequence of n random numbers and the time at which a k-peak occurs for the nth time in an infinite sequence of random numbers are asymptotically normally distributed. Exact expressions for their means and variances are evaluated.

IN his twenty-first convocation address 'Uncertainty, randomness and creativity' delivered at the Indian Statistical Institute, Calcutta on 5th March 1987, C. R. Rao[1] brings to our attention, in his characteristic style, an attractive feature of random numbers:

'It is an interesting property of random numbers that, like the Hindu concept of God, it is patternless and yet has all the patterns in it. That is, if we go on generating strictly random numbers we will encounter any given pattern sometime or other. Thus, if we go on tossing a coin, we should not be surprised if 1000 heads appear in successive tosses at some stage. So we have the proverbial monkey which, if allowed to type continuously, can produce the entire works of Shakespeare in a finite though a long period of time. . . .'

Continuing in this vein he deals with the following example:

'It is found that population sizes of a large variety of animals exhibit roughly a three-year cycle, i.e., the average time that elapses between two successive peak years of population size is about three years. (A peak year is defined as a year in which there are more animals than in the immediately preceding and immediately succeeding years.) The ubiquity of such a phenomenon led some to believe that perhaps a new law of nature has been uncovered. The belief was dealt a mortal blow when it was noted that if one plots random numbers at equidistant points, the average distance between peaks approaches 3 as the series of numbers gets large. In fact, such a property is easily demonstrable by using the fact that the probability of the middle number being larger than the others in a set of 3 random numbers is $\frac{1}{3}$. This gives an average distance of 3 years between the peaks.'

Here we elaborate on these comments by presenting a more detailed but elementary analysis of the frequency of occurrence of peaks (of arbitrary order) in a sequence $\{\xi_n\}$, $n = 1, 2, \ldots$ of independent and identically distributed random variables with a continuous distribution function $F(x)$. We may and do, without loss of generality, assume that ξ_n is uniformly distributed in the interval $[0, 1]$. This can always be achieved by considering the transformed sequence $\{F(\xi_n)\}$. We would consider this as a good pedagogical illustration of the individual ergodic theorem of Birkhoff and the central limit theorem for an m-dependent stationary sequence of random variables[2].

Say that a time point n is a k-peak for the sequence $\xi_1, \xi_2, \ldots$ if the event

$$E_n = \{\xi_n > \max(\xi_{n-k}, \xi_{n-k+1}, \ldots, \xi_{n-1}, \xi_{n+1}, \xi_{n+2}, \ldots, \xi_{n+k})\} \quad (1)$$

occurs. Clearly the probability of E_n is given by

$$P(E_n) = \int_0^1 x^{2k}\, dx = \frac{1}{2k+1}. \quad (2)$$

For $k = 1$, this is $\frac{1}{3}$ as mentioned earlier.

Denote by S the shift transformation defined by $S\xi_n = \xi_{n+1}$, $n = 1, 2, \ldots$. The distribution of the i.i.d sequence $\xi_1, \xi_2, \ldots$ is invariant and ergodic under S. For any event E denoted by 1_E its indicator which assumes the value 1 if E occurs and 0 otherwise. Then

$$\begin{aligned}\pi_{n,k} &= 1_{E_{k+1}} + 1_{E_{k+2}} + \ldots + 1_{E_{n+k}} \\ &= 1_{E_{k+1}} + 1_{E_{k+1}} oS + 1_{E_{k+1}} oS^2 + \ldots + 1_{E_{k+1}} oS^{n-1}\end{aligned} \quad (3)$$

is the number of k-peaks in the finite length sequence $\xi_1, \xi_2, \ldots, \xi_{n+2k}$. Note that the first likely k-peak in this sequence cannot be before the time $k+1$ and similarly the last likely k-peak cannot be after the time $n+k$. It now follows from the individual ergodic theorem of Birkhoff that

$$\lim_{n\to\infty} \frac{\pi_{n,k}}{n} = \frac{1}{2k+1} \text{ a.s.} \quad (4)$$

In other words the proportion of the number of k-peaks

K. R. Parthasarathy is at the Indian Statistical Institute, 7, S. J. S. Sansanwal Marg, New Delhi 110 016, India

in a sequence of $n+2k$ random numbers converges almost surely to $1/(2k+1)$ as $n \to \infty$.

We shall now evaluate the variance of $\pi_{n,k}$. First observe that $1_{E_{k+1}}$ and $1_{E_{k+r+1}}$ are independent random variables whenever $|s-r| \geq 2k$. The covariance between $1_{E_{k+1}}$ and $1_{E_{k+r+1}}$ depends only on $s-r$ (and k which is fixed). The variance of $1_{E_{k+1}}$ is

$$\frac{1}{2k+1}\left(1-\frac{1}{2k+1}\right) = 2k(2k+1)^{-2}.$$

Thus

$$\operatorname{var}(\pi_{n,k}) = \frac{2nk}{(2k+1)^2} + 2\sum_{1 \leq r \leq 2k}(n-r)\operatorname{cov}(1_{E_{k+1}}, 1_{E_{k+r+1}}), \quad (5)$$

where 'var' denotes variance, and 'cov' the covariance. If $1 \leq r \leq k$ it is clear that $E_{k+1} \cap E_{k+r+1} = \phi$.

Thus

$$\operatorname{cov}(1_{E_{k+1}}, 1_{E_{k+r+1}}) = -\frac{1}{(2k+1)^2} \quad \text{if} \quad 1 \leq r \leq k. \quad (6)$$

For $k < r \leq 2k$ we have

$$\begin{aligned}
&\mathbb{E}\, 1_{E_{k+1}} 1_{E_{k+r+1}} \\
&= P(\xi_{k+1} > \max(\xi_1, \xi_2, \ldots, \xi_k, \xi_{k+2}, \xi_{k+3}, \ldots, \xi_{2k+1}), \\
&\quad \xi_{k+r+1} > \max(\xi_{r+1}, \xi_{r+2}, \ldots, \xi_{r+k}, \xi_{r+k+2}, \ldots, \xi_{r+2k+1})), \\
&= \mathbb{E}\,(\xi_{k+1}\xi_{r+k+1})^{r-1}\,(\xi_{k+1} \wedge \xi_{k+r+1})^{2k+1-r} \\
&= 2\int_{0<x<y<1} x^{2k} y^{r-1}\, dx\, dy \\
&= \frac{2}{(2k+1)(2r+1)}, \quad (7)
\end{aligned}$$

where $\wedge$ denotes minimum. Substituting (6) and (7) in (5) we conclude that

$$\operatorname{var}(\pi_{n,k}) = \frac{-2nk}{(2k+1)^2} + \frac{2k}{2k+1} + \frac{4}{2k+1}\sum_{r=k+1}^{2k}\frac{n-r}{2k+r+1}. \quad (8)$$

This implies

$$\lim_{n\to\infty}\frac{1}{n}\operatorname{var}(\pi_{n,k}) = \frac{2}{(2k+1)^2}\sum_{r=k+1}^{2k}\frac{2k+1-r}{2k+1+r} = \sigma_k^2, \quad \text{say.} \quad (9)$$

The following table gives the values of σ_k, $1 \leq k \leq 5$

k	1	2	3	4	5
σ_k	0.21082	0.16997	0.14517	0.12858	0.11655

By the central limit theorem for an m-dependent stationary sequence of random variables (see for example Theorem 19.2.1, ref. 2) we now conclude that

$$\left[\left(\pi_{n,k} - \frac{n}{2k+1}\right)\Big/(\sqrt{n}\,\sigma_k)\right] \overset{\text{in law}}{\Rightarrow} N(0, 1), \quad (10)$$

as $n \to \infty$, where $\Rightarrow$ denotes weak convergence and $N(0, 1)$ denotes the standard normal distribution with density function $(2\pi)^{-1/2}\exp -\frac{1}{2}x^2$. It may be noted that for large k, σ_k is approximately

$$\left\{\frac{2}{2k+1}\left(2\log\tfrac{4}{3} - \tfrac{1}{3}\right)\right\}^{1/2} = 0.38824(2k+1)^{-1/2}.$$

Denote by $\tau_{m,k}$ the time at which the mth k-peak occurs. Then

$$\begin{aligned}
P(\tau_{m,k} \geq n) &= P(1_{E_{k+1}} + 1_{E_{k+2}} + \ldots + 1_{E_{k+n}} \leq m) \\
&= P(\pi_{n,k} \leq m) \\
&= P\left[\left(\pi_{n,k} - \frac{n}{2k+1}\right)\Big/\sqrt{n}\,\sigma_k \right. \\
&\leq \left.\left(m - \frac{n}{2k+1}\right)\Big/\sqrt{n}\,\sigma_k\right]. \quad (11)
\end{aligned}$$

Putting

$$\left(m - \frac{n}{2k+1}\right)\Big/\sqrt{n}\,\sigma_k = \alpha$$

and solving for n we get

$$n = (2k+1)^2/4\ \{2\alpha^2\sigma_k^2 + [4m/(2k+1)] - 2\alpha\sigma_k(\alpha^2\sigma_k^2 + [4m/(2k+1)])^{1/2}\}.$$

From (10) and (11) we now conclude that

$$\lim_{m\to\infty} P\left(\frac{\tau_{m,k} - m(2k+1)}{\sqrt{m}\,\sigma_k(2k+1)^{3/2}} > -\alpha\right) = \frac{1}{\sqrt{2\pi}}\int_{-\alpha}^{\infty} e^{-x^2/2}\, dx.$$

In other words

$$\sqrt{m}\,\frac{(\tau_{m,k}/m) - \overline{2k+1}}{\sigma_k(2k+1)^{1/2}} \overset{\text{in law}}{\Rightarrow} N(0, 1) \quad (12)$$

as $m \to \infty$. Roughly speaking the average length of the

time interval between the two successive k-peaks is $2k+1$. When $k=1$ this is 3 as pointed out in the beginning.

It is not difficult to see that the joint law of

$$\left[\sqrt{m} \left(\frac{\tau_{mj}}{m} - \overline{2j+1} \right),\ j = 1, 2, \ldots, k \right]$$

converges to a k-dimensional normal distribution with a nondiagonal covariance matrix as $m \to \infty$. The asymptotic results (10) and (12) provide a series of large sample tests for testing the hypothesis that a sequence of numbers is random.

1 Rao, C. R , Uncertainty, Randomness and Creativity, 21st convocation address, Indian Statistical Institute, Calcutta, 1987.

2. Ibragimov, I A. and Linnik, Yu. V., *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff Publishing, Groningen, 1971, pp 369–370

Biodiversity conservation information network: A concept plan

C. P. Geevan

A network that ensures the availability of reliable, up-to-date environmental information is necessary to realize the objectives set out in Convention on Biodiversity (now a treaty) that followed from the Earth Summit at Rio, June 1992. The task of building a nature conservation information network should, therefore, be considered an important part of the biodiversity conservation agenda. This paper presents an outline of an hypothetical information network, designated as Conservation Information Network (CiNet), to meet requirements of bioresources conservation, mapping, inventorying and monitoring on a large scale. The dataflow framework presented takes into account the existing data networks in India.

Background

The Convention on Biological Diversity (CBD) that emerged from the United Nations Conference on Environment and Development (UNCED) or the Earth Summit at Rio de Janeiro, in June 1992 is now a treaty. The CBD covers almost every aspect of biodiversity conservation. Article 17 of CBD concerns exchange of information. However, it does not lay down any operational framework for achieving information exchange. Nevertheless, an information network that ensures the availability of reliable, up-to-date environmental information is necessary to realize the objectives set out in CBD. The existing information systems are considered to be inadequate to meet these challenges[1].

The task of building a nature conservation information network should, therefore, be considered an important part of the conservation agenda. This paper presents an outline of an information network to meet requirements of bioresources mapping, inventorying and monitoring programme. The hypothetical network is designated as Conservation Information Network (CiNet). The dataflow framework is presented taking into account the ready availability of well-developed data networks in India and the CiNet is conceptualized as an overlay network riding over the existing networks.

C. P. Geevan is in the Salim Ali Centre for Ornithology and Natural History, Kalampalayam P O., Coimbatore 641 010, India.

Special interest groups on global networks

Computer networking has gone beyond setting up data links to creating information highways over which organizations and individuals are connected across the globe. The nature conservation efforts need to take advantage of these developments in information technology and create a niche for itself in the cyberspace. There are several important initiatives in this direction such as the INFOTERRA of the United Nations Environment Programme (UNEP) consisting of 170 national nodal points coordinated from the UNEP headquarters at Nairobi, the Environmental Resources Information Network (ERIN) in Australia with a biodiversity information system designed to meet the changing user needs, the Bio-diversity Information Network (BIN21) dedicated to the CBD with its secretariat at the Tropical Database