Yakov G. Sinai, Abel Prize Laureate 2014

ENTROPY OF 0,1-SEQUENCES

We consider a dynamical system where the state space consists of all infinite 0,1-sequences and where the dynamics is given by the shift operator. This dynamical system is named after the great seventeenth century mathematician, Jacob Bernoulli.

Consider the following 0,1-sequence made up of 50 digits:

11010010001010111011011 00
01010101000111001101000 11

Do we have any reason to believe that this sequence is randomly generated?

We state some relevant facts about the sequence.

1. The sequence contains 25 0's and equally many 1's. This fits with an assumption of randomness.

2. In 30 positions the sequence switches from 0 to 1 or vice versa, leaving 19 positions where the next digit is the same as the previous one. In a random sequence these numbers would tend to be the same.

3. The sequence contains six subsequences with 3 consecutive digits being equal, but none with 4. In a randomly generated sequence of 50 digits, the probability of finding a subsequence of at least 4 consecutive digits being equal is approxiomately 98 %. The fact that our sequence has no such subsequence indicates that it is not randomly generated.

Based on these arguments we conclude that we do not believe that the sequence is randomly generated.

The truth is that the sequence is manually generated in an attempt to produce a sequence which looks random. Our mistake, as our small analysis shows, is that we have switched digits to often.

Entropy of Bernoulli schemes

A 0,1-sequence is known as a Bernoulli scheme. In a randomly generated process, we have equal probability $p = \frac{1}{2}$ for the digits 0 and 1. In our example it seems that the occurences of 0 and 1 have equal probability, but that the combinations 01 and 10 are more likely, say 60/40 %, than the combinations 00 and 11. This difference in the predictability is quantifieded in the **entropy** of the system. The more unpredictable a system is the higher is the entropy. A random 0,1-sequence has entropy 0,693. Our example has entropy 0,673, which is slightly lower. In general, a Bernoulli scheme of two outcomes of probability p and $1-p$ has entropy given by

$$E = -p\ln p - (1-p)\ln(1-p)$$

A Bernoulli scheme may have more than two outcomes. The set of all infinite sequences of letters is a Bernoulli scheme of 26 outcomes. The famous mathematician John von Neumann asked an intriguing question about Bernoulli schemes. He wondered if it is possible that two structurally different Bernoulli schemes can produce the same result. Is it possible to identify the two Bernoulli schemes $BS(\frac{1}{2},\frac{1}{2})$ and $BS(\frac{1}{3},\frac{1}{3},\frac{1}{3})$? BS stands for Bernoulli scheme and the fractions give the probability for each outcome, thus $BS(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ is the Bernoulli scheme of three outcomes of equal probability.

The solution to the question of von Neumann was finally given by Donald Ornstein in 1970. The answer was no, two essentially different Bernoulli schemes provides different results. The basis for this result was given by Sinai and Kolmogorov i 1959. It turns out that the Kolmogorov-Sinai-entropy is precisely what separates different Bernoulli schemes.

A combinatorial result

In the beginning of this article it was claimed that the probability that a random 0,1-sequence of 50 digits contains a subsequence of 4 consecutive digits being equal is 98 %. We shall consider a slightly more general version of this claim.

A 0,1-sequence of length n is chosen randomly. What is the probability that the sequence contains m consecutive digits being equal, either 0 or 1?

The answer is given recursively. Let X_n be the set of 0,1-sequences of length n. The cardinality of X_n is obviously 2^n. Let $q(n,m)$ be the number of sequences containing m consecutive digits being equal, either 0 or 1, and $p(n,m)$ be the probability of picking such a sequence, i.e. $p(n,m) \cdot 2^n = q(n,m)$.

Let $\xi \in X_{n+1}$ be a sequence, containing m consecutive digits being equal. Remove the last digit from the sequence. Then the new sequence:

1. Contains a sequence of m consecutive digits being equal, or

2. Ends with a sequence of precisely $m-1$ consecutive digits being equal, i.e. the preceding digit of the last $m-1$ digits is different from the subsequent, or

3. Both 1. and 2.

The cardinality of the three sets are as follows. The first one has $q(n,m)$ elements, the second one has 2^{n+1-m} elements, and the third has $2^{n+1-m} \cdot p(n+1-m,m)$. This gives the recurrence relation

$$q(n+1,m) = 2q(n,m) + 2^{n+1-m} - 2^{n+1-m} \cdot p(n+1-m,m)$$

or for the probabilities, by dividing through by 2^{n+1},

$$p(n+1,m) = p(n,m) + 2^{-m}(1 - p(n+1-m,m))$$

It is more or less obvious that $p(n,1) = 1$, that $p(n,n) = 2^{1-n}$ and that $p(n,m) = 0$ for $n < m$. The following table shows some computations:

n/m	2	3	4	5	6
1	0	0	0	0	0
5	0,938	0,5	0,188	0,063	0
10	0,998	0,826	0,465	0,217	0,094
20	1	0,979	0,768	0,458	0,237
35	1	0,999	0,934	0,689	0,410
50	1	1	0,981	0,821	0,544

Observe that for sequences of 50 digits, there is a majority of sequences contaning a subsequence of 6 consecutive equal digits, and in fact 98 % of the sequences contain a subsequence of 4 consecutive equal digits.