UNSOLVED PROBLEMS

Edited by **Richard Guy**

In this department the MONTHLY *presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics and Statistics, The University of Calgary, Alberta. Canada T2N 1N4.*

A Pseudorandom Sequence—How Random Is It?

Andrzej Ehrenfeucht and Jan Mycielski

Let $\varepsilon_1, \varepsilon_2, \dots$ be a sequence of 0's and 1's. Suppose that we know $\varepsilon_1, \dots, \varepsilon_n$ and are asked to predict ε_{n+1}. A very simple way, which we will call the method M, is the following. Find the longest final segment $\varepsilon_j, \varepsilon_{j+1}, \dots, \varepsilon_n$ which occurs earlier in $\varepsilon_1, \dots, \varepsilon_n$. So $n - j$ is maximal such that $(\varepsilon_j, \varepsilon_{j+1}, \dots, \varepsilon_n) = (\varepsilon_{j-i}, \varepsilon_{j+i+1}, \dots, \varepsilon_{n-i})$ for some $i > 0$. Then find the smallest i (the most recent occurrence) for which this is so and let ε_{n-i+1} be your guess for ε_{n+1}. (Note that if $(\varepsilon_1, \dots, \varepsilon_n) = (\varepsilon, \varepsilon, \dots, \varepsilon, 1 - \varepsilon)$, then $(\varepsilon_j, \dots, \varepsilon_n)$ is empty and $i = 1$. Otherwise $(\varepsilon_j, \dots, \varepsilon_n)$ has length ≥ 1). The method M may seem to be very naive, but more or less refined variants of this method are used by all learning organisms. Perhaps every sensible method of prediction based on experience is equivalent to some kind of coding or description of the past by means of a sequence of 0's and 1's and the method M. Notice that if the sequence $\varepsilon_1, \varepsilon_2, \dots$ is eventually periodic, the predictions by M are eventually faultless.

In this note we do not consider any coding and use M only to produce a certain pseudorandom sequence $\rho_1, \rho_2, \dots$ We put $\rho_1 = 0$ and assume that whenever M predicts ρ_{n+1} to be ε, then in fact $\rho_{n+1} = 1 - \varepsilon$. Thus $\rho_1, \rho_2, \dots$ is characterized by the assumptions that $\rho_1 = 0$ and that M is always wrong. We could say that, from the point of view of M, the sequence $\rho_1, \rho_2, \dots$ is the most unpredictable one. It is easy to find by hand the first 40 values of this sequence:

$$\begin{aligned}(\rho_1, \rho_2, \dots) = (&0,1,0,0,1,1,0,1,0,1,1,1,0,0,0,1,0,0,0,0,1,1,\\ &1,1,0,1,1,0,0,1,0,1,0,0,1,0,0,1,1,1,\dots).\end{aligned}$$

Theorem. *Every finite sequence of* 0'*s and* 1'*s occurs as a segment in* $\rho_1, \rho_2, \dots$.

Proof: Assume that this theorem fails. Then there exists a finite sequence which does not occur infinitely many times as a segment of $\rho_1, \rho_2, \dots$. Let $\varepsilon_1, \dots, \varepsilon_k$ be any such sequence which is the shortest. Then let S be the set of all left extensions of $\varepsilon_1, \ldots, \varepsilon_k$, that is sequences of the form $\eta_1, \ldots, \eta_r, \varepsilon_1, \ldots, \varepsilon_k$, which occur in $\rho_1, \rho_2, \ldots$. So, of course, S is finite. Since $\varepsilon_1, \ldots, \varepsilon_{k-1}$ occurs infinitely many times in $\rho_1, \rho_2, \ldots,$ there exists a sequence of the form $\eta_1, \ldots, \eta_s, \varepsilon_1, \ldots, \varepsilon_{k-1}$ which occurs infinitely many times in $\rho_1, \rho_2, \ldots$ and is longer than any sequence in S. Of course, $\eta_1, \ldots, \eta_s, \varepsilon_1, \ldots, \varepsilon_k$ does not occur at all in $\rho_1, \rho_2, \ldots$. Let $\rho_j, \rho_{j+1}, \ldots, \rho_{j+s+k-2}$ and $\rho_{j-i}, \rho_{j-i+1}, \ldots, \rho_{j+i+s+k-2}$ be the first two occurrences of $\eta_1, \ldots, \eta_s, \varepsilon_1, \ldots, \varepsilon_{k-1}$ in $\rho_1, \rho_2, \ldots$. Since the method M never predicts correctly any ρ_n, it does not predict correctly $\rho_{j+s+k-1}$. Hence $\rho_{j+s+k-1} \neq \rho_{j-i+s+k-1}$. Therefore, either $\rho_j, \ldots, \rho_{j+s+k-1}$ or $\rho_{j-i}, \ldots, \rho_{j-i+s+k-1}$ equals $\eta_1, \ldots, \eta_s, \varepsilon_1, \ldots, \varepsilon_k$, which is a contradiction. So the theorem is proved.

Remark. The above theorem remains true if we modify the definition of $\rho_1, \rho_2, \ldots$ initiating it with any finite sequence of 0's and 1's.

Now our problem is how random is the sequence $\rho_1, \rho_2, \ldots$? And the same question can be raised about the modifications mentioned in the remark. Of course, from an algorithmic point of view, they are not random at all since there exist programs for producing them. But, from a statistical point of view, they could be quite random. For example, do they satisfy

$$\frac{\rho_1 + \cdots + \rho_n}{n} \to \frac{1}{2}?$$

The first 1300 values of the sequence, calculated by Walter Taylor.

```
0 1 0 0 1 1 0 1 0 1 1 1 0 0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 0 1 0 1 0 0 1
0 0 1 1 1 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 0 1 1
1 1 1 1 0 1 0 1 0 1 1 0 0 0 1 1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1
1 0 1 1 0 1 0 0 1 1 0 0 0 1 0 1 1 1 1 0 0 0 0 1 0 0 1 0 1 1 0 0 1 1 0
1 1 1 0 1 1 1 1 0 1 0 0 1 0 1 0 1 1 1 1 1 1 1 0 0 0 1 1 0 1 0 0 0 0 1
0 1 0 0 0 1 0 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 1 1 0 1
1 0 1 1 0 1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1 1 0 1
0 1 1 0 1 0 1 0 0 1 1 1 1 0 0 1 0 1 1 1 0 1 1 0 0 0 1 0 0 1 0 0 1 0 1
0 0 0 0 1 1 0 1 0 1 0 1 0 1 1 1 0 1 0 0 1 1 1 0 0 1 1 0 1 0 0 1 0 0 0
1 1 0 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 0 1 0 0 1 1 0
0 1 0 1 1 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 1 1 1 0
1 1 1 0 0 0 0 1 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 0 0 1 1 0 1 0 0 0 1 1 1
0 1 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1 1 1
1 1 1 0 1 1 0 1 0 1 1 1 1 1 0 1 0 0 0 0 0 1 0 0 1 1 1 1 1 0 0 0 0 0 1
1 0 0 0 1 1 0 0 1 1 0 0 0 0 1 0 1 1 1 0 0 1 0 1 0 1 0 1 0 0 1 0 1 1 1
1 1 1 0 1 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 0 1 1 0 0 0 0 0 1 1 1 1 0 0 1
1 0 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 1 0 0 1 1 1 1 0 1 1 1 0 0 1 0 0 1 0
1 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 1 0 0 1 0 1 1 1 1 0 1 0 1 0 0 1 0 0 0
0 1 1 0 0 0 0 1 1 1 1 1 0 0 1 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 0 0 0 1 0
0 1 0 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0
0 1 0 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 0 1 0 1 1 0 1 0 1 1 0 1 1 1 0 1
0 0 0 0 1 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 1 0 1 0 1 1 1 1 0 0 1 0 0 0 0
0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 0 0 1 1
0 0 1 0 0 0 1 1 1 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 1 0 1 0
1 0 1 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 1 1 0 1 0 0 1 1 0 1
1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 1 0 0 0 0 0 0 0
1 1 0 0 1 1 1 0 0 1 0 1 1 0 0 0 1 0 1 0 1 1 1 0 0 1 1 0 0 0 1 1 1 1 1
1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 1 0 1 0 1 1 0 0
1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 1 1 1 1 0 1 0 1 1 0 0 0 0 1 0 0 0 1 0
1 1 1 0 1 0 1 0 1 0 1 0 1 1 0 0 1 0 1 0 1 0 0 0 1 1 1 0 0 1 1 0 0 1 0
1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 1 1 0 0 0 1 0 0 0 1 0 1 0 0 1 0 0
0 1 0 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 1 0 1 1 0 0 1 1 0 0 0 1 0 0 1 1
0 0 1 1 1 0 1 0 0 1 1 0 0 1 0 0 1 1 0 1 1 1 0 0 1 0 1 1 1 1 1 0 1 1 1
1 1 0 0 0 0 1 0 1 1 0 1 0 0 1 1 1 0 1 0 1 0 1 1 1 0 0 0 0 1 0 0 0 0 1
1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 0 1
1 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1 1 1 1 0 1 0 1 0 1 0 0 1 1 1 0 0 0 0
1 1 0 0 0 1 0 1 0 0 0 1 1 0 0 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1
1 1 0 1 1
```

Further comments by I. J. Good. A Mycielski sequence can be expected to be flatter than "flat-random" because it is constructed to avoid repeated subsequences to some extent. An appropriate test for this purpose, over *finite* stretches, would be the *serial test*, the correct use of which is explained by Good (1953) and exemplified for the binary expansion of $\sqrt{2}$ by Good and Gover (1967). Since Walter Taylor has already written a program for generating M-sequences it would be easy for him to apply the serial test, and he will presumably thereby corroborate my expectation. Note, however, that the further one goes in the sequence the more one is avoiding longer repeats so the Mycielski sequence is not homogeneous. Meanwhile, I counted by hand the numbers of 1s in each of the 37 rows of length 35 in the printout and obtained a Pearson chi-squared value of only 15.7 with 36 degrees of freedom, corresponding to a P-value of 0.9987 (assuming the asymptotic chi-squared distribution). This supports my conjecture over the first 1295 bits.

A Mycielski sequence could also be called a Gambler's Fallacy sequence. Another class of Gambler's Fallacy sequences can be defined recursively in the following manner: at each stage of the construction choose a digit that will provide a new polybit of length k (a k-bit) where, at that stage, k is small as possible. When this rule does not determine whether a 0 or a 1 should be the next bit, decide by tossing a coin (or by a deterministic rule is preferred). Here is an example: $01\overset{*}{0}0\overset{*}{1}1\overset{*}{1}01\overset{*}{0}11000\overset{*}{0}10\ldots$ where the asterisks indicate the bits that had to be chosen at random. Presumably such a sequence is even more flatter-than-random than a Mycielski sequence.

REFERENCES

1. I. J. Good, The serial test for sampling numbers and other tests for randomness," *Proc. Cam. Philos. Soc.* 49, (1953) 276–284.
2. I. J. Good and T. N. Gover, The generalized serial test and the binary expansion of $\sqrt{2}$, *J. Roy. Statist. Soc. A 130*, (1967) 102–107; 131 (1968), 434.

Department of Computer Science
University of Colorado
Boulder, CO 80309

Computer Research and Applications Group
Los Alamos National Laboratory
Los Alamos, NM 87545