of these select the $(t+1)$st element.

The idea developed in the preceding paragraph leads immediately to the following algorithm:

Algorithm S (*Selection sampling technique*). To select n records at random from a set of N, where $0 < n \leq N$.

S1. [Initialize.] Set $t \leftarrow 0$, $m \leftarrow 0$. (During this algorithm, m represents the number of records selected so far, and t is the total number of input records we have dealt with.)

S2. [Generate U.] Generate a random number U, uniformly distributed between zero and one.

S3. [Test.] If $(N-t)U \geq n-m$, go to step S5.

S4. [Select.] Select the next record for the sample, and increase m and t by 1. If $m < n$, go to step S2; otherwise the sample is complete and the algorithm terminates.

S5. [Skip.] Skip the next record (do not include it in the sample), increase t by 1, and go to step S2. ▮

This algorithm may appear to be unreliable at first glance and, in fact, to be incorrect; but a careful analysis (see the exercises below) shows that it is completely trustworthy. It is not difficult to verify that

a) At most N records are input (we never run off the end of the file before choosing n items).

b) The sample is completely unbiased; in particular, the probability that any given element is selected, e.g., the last element of the file, is n/N.

Statement (b) is true in spite of the fact that we are *not* selecting the $(t+1)$st item with probability n/N, we select it with the probability in Eq. (1)! This has caused some confusion in the published literature. Can the reader explain this seeming contradiction?

(*Note:* When using Algorithm S, one should be careful to use a different source of random numbers U each time the program is run, to avoid connections between the samples obtained on different days. This can be done, for example, by choosing a different value of X_0 for the linear congruential method each time; X_0 could be set to the current date, or to the last X value generated on the previous run of the program.)

We will usually not have to pass over all N records; in fact, since (b) above says that the last record is selected with probability n/N, we will terminate the algorithm *before* considering the last record exactly $(1-n/N)$ of the time. The average number of records considered when $n=2$ is about $\frac{2}{3}N$, and the general formulas are given in exercises 5 and 6.

Algorithm S and a number of other sampling techniques are discussed in a paper by C. T. Fan, Mervin E. Muller, and Ivan Rezucha, *J. Amer. Stat. Assoc.* **57** (1962), 387–402. The method was independently discovered by T. G. Jones, *CACM* **5** (1962), 343.