Def Sci J, Vol 37, No 3, July 1987, pp 361-366.

Generating Random Sampling Numbers

Chatter Singh

DRDO Computer Centre, Delhi–110 054

ABSTRACT

Simulation experiments with the help of random numbers are increasing in various Defence application areas. Some authors have brought out autocorrelations in the sequences generated by the computer methods. To overcome this defect, a new algorithm based on the concept of random sampling (with replacement) from a finite population has been developed. The constants in the linear congruential relation are chosen so that (a) the proportion of distinct numbers in the generation compares with the theoretical expectation and (b) the period of the generator covers almost the entire population of 16–bit numbers. The criteria suggest to choose the variable multiplier ($a_i = 2a_{i-1} + 1$) with initial value $a_0 = 2^9 + 5$ or $2^{10} + 3$. The randomness of the generation has been compared with that of a Bernoullian sequence.

1.INTRODUCTION

The first method for generating random numbers on a digital computer was mid-square method. The method was difficult to analyze, comparatively slow and poor in statistical properties and, therefore, was abondoned in favour of congruential methods. At present most procedures in use, are based on the congruential relationship defined by,

$$x \equiv (ax_{i-1} + c)\ (mod\ m), \quad i=1,2,3,......(1)$$

where x_0, a, c and m are suitably chosen integers and x transformed by x /m to approximate a $U(0,1)$ process. Random samples from any other distribution are obtained by transformations. For some distributions, the transformations from uniform are simple and exact; in other cases the more complicated transformations are

Received 12 December 1985

approximated. Sowey[1] cited about 450 contributions in the bibliography on Random number generation and testing. Some authors have brought out serious short-comings in the methods for generating random numbers. Marsaglia[2] describes the situation that the random numbers fall mainly in the planes. Kennedy and Gentle[3] reviewed the various methods critically but find it difficult to choose an *a* guaranteeing good 'random' sequence and believe that given a generator one or more tests could be formed that the generator would fall. Atkinson[4] points out unsuitably of the spectral and lattice tests for generators other than full-period.

Here, we consider to simulate a random sampling process (with replacement) from a finite population. The process repeats some numbers and covers the population probabilistically as the sample size increases. This property of randomness guides to choose *a* and *c* in Eqn (1) so that the number of distinct numbers compares with the expectation for any sample size *n*, and the period of the generation covers almost the entire population numbers. Theoretical properties of the sampling process are derived and the choice of the generator with comparable properties made. Randomness of the generation has then been compared with that of Bernoullian sequence.

2. CHARACTERISTICS OF THE RANDOM PROCESS

Consider a sample of size *n* drawn (with equal probabilities and replacement) from a population numbered from 1 to *N*. Let *S* be the number of distinct number appearing in the sample. Following Feller[5] and Basu[6], it is found that

$$P(S=s) = N^{-n} \binom{N}{s} \left[s^n - \binom{S}{1}(s-1)^n + \binom{S}{2}(s-2)^n - \ldots\ldots\right] \qquad (2)$$

where *s* varies from 1 to smaller of *n* and *N*.

(Leading term s^n reported by Basu as 1 seems to be in error).

In terms of the 'differences of zeros', we may write the Eqn (2) in the form

$$P(S=s) = N^{-n} \binom{N}{s} \Delta^s\, 0^n \qquad (3)$$

where Δ is usual difference operator with unit increments and $\Delta^s 0^n$ interpreted as $\Delta^s x^n$ at $x = 0$.

From Eqn (3), the probability generating function of *s*

$$P(t) = N^{-n} \sum_{s=0}^{N} \binom{N}{s} t^s \Delta^s 0^n = N^{-n} (1 + t\Delta)^N 0^n \qquad (4)$$

and the factorial moment generating function is

$$F(t) = N^{-n} (E + t\Delta)^N 0^n \qquad (5)$$

where $E = \quad + \Delta$

Whence

$$E(S) = N^{-n} \binom{N}{1} E^{N-1} \Delta^s 0^n$$
$$= N^{-n} N[E^N - E^{N-1}]\, 0^n$$

$$N\,[1 \qquad \frac{N-1}{N}\Big)^{\prime}$$

For large N and $\lambda = n/N$

$$E(S) = N\,(1 - e^{-\lambda})$$

Similarly

$$V(S) = N \left(\frac{N-1}{N}\right)^{n} - N^{2} \left(\frac{N-1}{N}\right)^{2n} + N\,(N-1) \left(\frac{N-2}{N}\right)^{n}$$

$$= N\,e^{-\lambda}\,(1 - e^{-\lambda})$$

For $N=2^{15}$ and $\lambda = 1(1)6$, the expectation and standard error of S/N are given in Table 1

Table 1. Expectation and standard error of *(S/N)*

(expressed in units of 10^{-4})

	1	2	3	4	5	6	7
E(S/N)	6321	8646	9502	9816	9932	9975	9990
SE(S/N)	26	18	12	7	4	2	1

This suggests that the random sampling as large as the population size limits to 63 percent distinct numbers of the population. In order that all possible samples in the population are accessible to a generator, the period need to be at least five times the population size.

3. ALGORITHM

Consider a generator repeating the following Fortran instructions in a 16 bit word

```
      IA = 2*IA + 1
      IX = IA*1X + 1
      IF (IX.GT.0) GO TO 5
      IX = IX + 32767 + 1
      IA = 32767 - IA
    5 CONTINUE
```

(6)

Let $IA = IX = 2^{p} + j$, (p=4,5,   14; j=1,3,   11)

The procedure provides distinct numbers close to the expectation till its period which is found atleast N in all the cases excepting $(2^{5} + 11)$. Comparison of sequence of lengths, $n=N$, $3N$ & $5N$ has been reported in Table 2. It may be observed that longer periods are attained by the multipliers $2^{9}+5$, $2^{10}+3$ and $2^{11}+3$. However, it is interesting to note that these sequences starting at different points sooner or later join the same path before repeating themselves.

Table 2. Values, (S – E (S))/N, in the sequence length *n*, produced by the multiplier, $2^p + j$

(Deviations expressed in units of 10^{-4})

n=	N					$3N$					$5N$				
$p \backslash j$=	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
4	–29	–37	4	–19	21	x	x	x	x	x	x	x	x	x	x
5	–2	36	32	–8	26	x	x	x	x	x	x	x	x	x	x
6	–11	10	–17	–11	–39	x	11	x	–18	x	x	–90	x	x	x
7	1	–12	–6	–3	–31	x	x	x	13	x	x	x	x	x	x
8	2	12	21	–22	–6	x	x	x	x	x	x	x	x	x	x
9	21	–3	18	1	–12	–30	x	13	x	x	x	x	–11	x	x
10	–18	–7	5	–5	20	5	12	x	x	x	x	13	x	x	x
11	–15	18	–1	–8	–11	–1	8	x	x	–13	x	–28	x	x	x
12	4	6	7	28	–45	x	x	13	–9	–3	x	x	–84	x	x
13	–23	–43	1	11	40	x	x	x	x	–10	x	x	x	x	x
14	–8	11	23	3	22	6	x	x	18	x	x	x	x	–77	x

* Sequence repeating

4. TESTS OF RANDOMNESS

The object of testing generated sequences is to verify that they have the particular random like qualities for a specified purpose. The sequence to produce distinct/repeated numbers comparable to the theoretical random pattern have been verified. This consideration permits sequences with longer periods. The sequences produced by the multipliers 2^9+5 and 2^{10}+3 are further compared with Bernoulli trials. In that n successive numbers are considered as a result of n independent trials and a success considered if the number is greater than 0.5m. The distribution of successes in N trials each of length n, is then compared with the binomial frequencies

$$N\binom{n}{r}(0.5)^r(0.5)^{n-r} \qquad r=0,1,\ldots..n \tag{7}$$

Comparisons (observed–expected–frequencies) for the multipliers 2^9+5 and 2^{10}+3 are given in Table 3. The χ^2 values suggest no dependence between the numbers of the sequence.

Table 3. Comparing the distribution of successes in 10,000 consecutive sets each containing n numbers produced by (a) 2^9+5 and (b) $2^{10}+3$, with the binomial frequencies.

		(a)			(b)	
n	5	10	15	5	10	15
		(observed – expected)				
0	–20	5	0	–19	6	0
	53	0	2	31	0	2
2	–5	21	–1	26	14	–1
	–47	38	13	54	52	13
	22	–12	45	74	–30	41
5	–1	–6	4	4	4	16
6		–43	11		–37	6
		–27	–24		–41	–33
		1	–61		8	–46
9		18	8		19	–11
		4	22		4	26
11			–13			–8
			–4			–3
13			–3			–3
			0			0
15			1			1
	4.11	9.65	13.09	3.05	12.73	9.53

5. CONCLUSIONS

It seems that the method provides a sample that satisfies the definition of the random sampling. The algorithm operating in 16-bit word and with no storage requirement, further, satisfies the criteria of a good algorithm. It may be noted that only two instructions are added to the poor generator RANDU. The importance of the method becomes greater as more and more minicomputers with smaller word lengths are being used.

REFERENCES

1. Sowey, E.R., *Inst. Statist. Rev.*, **40** (1978), 355-371.
2. Marsaglia, G., *Proc. Nat. Acad. Sci.*, **61** (1968), 25-28.

3. Kennedy, W.J & Gentle, J.E., Statistical Computing (Marcel Dekker, New York), 1978.

4 Atkinson, A.C., *Appl. Statist.*, **29** (1980), 164-171.

5. Feller, W., Introduction to Probability Theory and its Applications, (John Wiley, New York), Vol. 1, 1968.

6. Basu, D., *Sankya,* **20** (1958), 287-294.