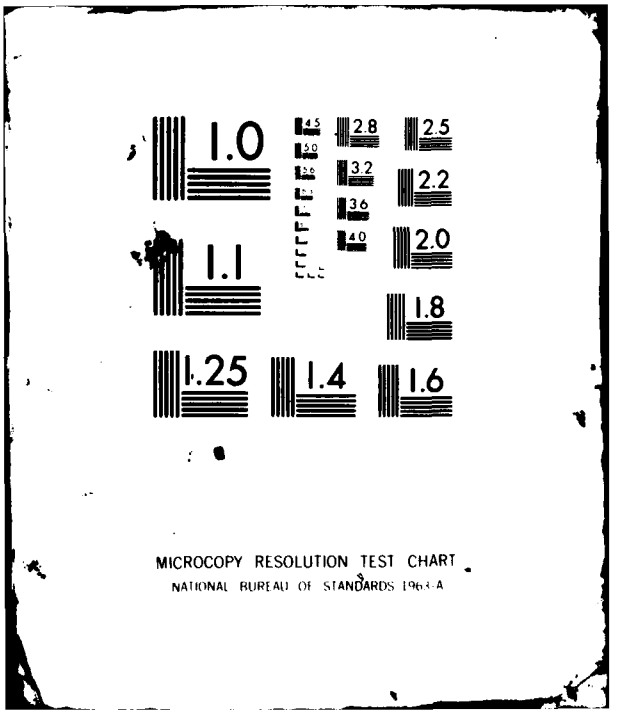
AD-A111 967 NORTH CAROLINA UNIV. AT CHAPEL HILL CURRICULUM IN OPER--ETC F/G 12/1
SAMPLING FROM A DISCRETE DISTRIBUTION WHILE PRESERVING MONOTONI--ETC(U)
FEB 82 G S FISHMAN, L R MOORE N00014-76-C-0302
UNCLASSIFIED UNC/ORSA/TR-81/7 NL

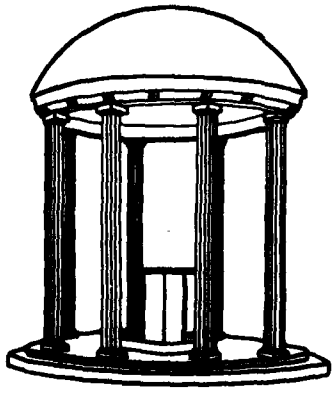
END
DATE FILMED
4 82
DTIC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

(12)

OPERATIONS RESEARCH AND SYSTEMS ANALYSIS

UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL



(12)

SAMPLING FROM A DISCRETE DISTRIBUTION

WHILE PRESERVING MONOTONICITY

George S. Fishman and Louis R. Moore, III

Technical Report No. UNC/ORSA/TR-81/7

Curriculum in Operations Research

and Systems Analysis

University of North Carolina at Chapel Hill

This research was supported by the Office of Naval Research under contract N00014-26-C-0302. Reproduction in whole or part is permitted for any purpose of the United States government.

ABSTRACT

This paper describes a cutpoint method for sampling from an n-point discrete distribution that preserves the monotone relationship between a uniform deviate and the random variate it generates. This property is useful when developing a sampling plan to reduce variance in a Monte Carlo or simulation study. The alias sampling method generally lacks this property and requires 2n storage locations while the proposed cutpoint sampling method requires m+n storage locations, where m denotes the number of cutpoints. The expected number of comparisons with this method is derived and shown to be bounded above by $(m + n - 1)/n$. The paper describes an algorithm to implement the proposed method as well as two modifications for cases in which n is large and possibly infinite.

I. INTRODUCTION

Let X be a discrete random variable on the integers $1,\ldots,n$ with probability mass function $\{p_i;\ i=1,\ldots,n\}$ and distribution function (d.f.)

$$q_0 = 0$$
$$q_i = q_{i-1} + p_i \qquad i=1,\ldots,n. \tag{1.1}$$

One straightforward way to sample from $\{q_i;\ i=1,\ldots,n\}$ is to sample U from the uniform distribution on $[0,1)$ and then determine X from

$$X = \min\{i:\ q_i \geq U\}. \tag{1.2}$$

If the d.f. of X is stored in a table beforehand, this procedure, known as the inverse transform method, requires n storage spaces and EX comparisons on average, which may prove costly when n is moderate or large.

At present, the most efficient way to sample repeatedly from the d.f. of X is the alias method of Walker (1974a, 1974b, 1977). See also Kronmal and Peterson (1979). This method requires storage for two arrays, the aliases $A_1,\ldots,A_n$ and alias probabilities $\{r_i = pr(X=i|L=i);\ i=1,\ldots,n\}$ where for $i=1,\ldots,n$

$$pr(L=i) = 1/n$$

$$pr(X=i|L=i) + pr(X=A_i|L=i) = 1 \qquad (1.3)$$

and

$$p_i = pr(X=i) = \frac{1}{n} \sum_{j=1}^{n} pr(X=i|L=j).$$

Prior to beginning the sampling experiment one chooses these arrays to satisfy (1.3). After sampling U one computes $L = \lfloor nU \rfloor + 1$ and selects $X = L$ if $r_L \geq U \pmod{1/n}$ or otherwise selects $X = A_L$. Here $\lfloor \theta \rfloor$ denotes the largest integer less than or equal to θ. Note that only one comparison is required to generate each X. The arrays determined by (1.3) require 2n storage locations and are not unique. If one wishes to retain the tabled values of $\{q_i\}$, an additional n storage locations are required.

Although the time independent nature of the alias method has clear appeal, the method has two limitations that deserve attention:

a. In general, the alias method does not preserve a monotone relationship between U and X as does the inverse transform method (1.2).

b. The allocation of 2n storage spaces may be infeasible either due to the magnitude of n or the requirements of other steps in the program in which the alias method is imbedded. Furthermore, if one wishes to maintain the table of $\{q_i\}$, 3n storage locations are required.

While the issue of storage requirements is self explanatory, the significance of the monotone property needs clarification. Let Y_1 and Y_2 be random variables with d.fs. F_1 and F_2 and inverse d.fs. $G_1(u) = \min(y: F_1(y) \geq u)$ and $G_2(u) = \min(y: F_2(y) \geq u)$ respectively. Then the minimal correlation between Y_1 and Y_2 occurs for $Y_1 = G_1(U)$ and $Y_2 = G_2(1-U)$. The result is due to Hoeffding (1940). See also Whitt (1976). In Monte Carlo sampling and discrete event simulation one often wants to make use of the minimal correlation property to induce a variance reduction for a given sampling cost. More generally, one often can achieve a variance reduction by appropriate use of the sequence $U_k = U + (k-1)/r \pmod 1$ for $k=1,\ldots,r$ with a sampling technique that preserves monotonicity. See Hammersley and Handscomb (1964) and Fishman and Huang (1980). The alias method may prevent one from effecting this reduction in variance.

2. THE CUTPOINT METHOD

We now describe the cutpoint method for sampling from the d.f. of X. The procedure preserves monotonicity, maintains the table of $\{q_i\}$ and allows the user to adjust space and time requirements to accommodate the global needs of the problem setting. The procedure again uses the inverse transform approach but with more information computed beforehand, as in the alias method. The proposed method is not new having been described in Chen and Asau (1974). However, the present paper is the first to study the tradeoff between computation time and space analytically.

For a given positive integer m, define the cutpoints

$$I_k = \min \{i: q_i > (k-1)/m\} \quad k=1,\ldots,m \qquad (2.1)$$
$$I_{m+1} = n.$$

Let $L = \lceil mU \rceil$ so that

$$pr(I_L \leq X \leq I_{L+1}) = 1 \qquad (2.2)$$

where X is as defined in (1.2) and $\lceil \theta \rceil$ denotes the smallest integer greater than or equal to θ. The maximal number of comparisons needed to determine X exactly is $I_{L+1} - I_L + 1$ and the expected maximal number of comparisons is

$$C_{m,n} = (I_{m+1} - I_1 + m)/m. \qquad (2.3)$$

The storage requirements are m+n locations with the tabled d.f. of X comprising n of them.

Note that $C_{n,n}$ is less than 2, implying that the cutpoint method requires less than one additional comparison on average to preserve monotonicity when compared to the alias method with the same allocation of storage. If the d.f. table of X also is to be maintained in the alias method, then for equal storage for the cutpoint method $C_{2n,n}$ is less than 1.5 so that at most 1/2 of an additional comparison is needed on average.

A more revealing evaluation arises if the d.f. of X is directly taken into account. Let J_m denote the number of comparisons on a trial. Then J_m equals $X - I_L + 1$, where X is determined as in (1.2), and has expectation

$$EJ_m = 1 + EX - EI_L$$
$$= 1 + EX - \sum_{\ell=1}^{m} I_\ell/m \quad . \tag{2.4}$$

Table 1 illustrates the proposed method using the eight-point distribution in Fishman (1978, p. 459).

Insert Table 1 about here

If space is not an issue then the alias method is the procedure of choice. One then may view $EJ_m - 1$ as the cost of maintaining the monotone property.

3. THE CASE OF LARGE n

If $\{q_i\}$ has infinite support ($n = \infty$) then neither the alias method nor our cutpoint method alone suffices to perform sampling. This insufficiency also may occur if n is merely large relative to space availability. Here Kronmal and Peterson (1979) suggest using the alias method "for a finite (but large) range of the desired discrete distribution and a special tail-generating method for the tail beyond". Ahrens and Dieter (1973) discuss the tail-generating methods for several common parametric families of distributions. More recently Ahrens and Kohrt (1981) described a cutpoint method with a more dense frequency of cutpoints in the tails.

Our own proposals for this situation take two forms. Note that in principle (2.4) suggests that EJ_m may be determined for infinite n if EX is finite. Consider $n^* < n$ such that a procedure is available for computing $\{q_i;\ i > n^*\}$. The cutpoint

method then applies directly with tables used for $\{q_i;\ i \leq n^*\}$ and the available procedure for $i > n^*$. The mean number of comparisons remains the same while the mean cost in time is proportional to

$$\mu_1 = EJ_m + c_1\, E(J_m | X > n^*)\,pr(X > n^*) \qquad (3.1)$$

where c_1 is the (computer dependent) increase in time required to evaluate q_i for $i > n^*$ relative to the time required for a table lookup of q_i when $i \leq n^*$.

As an alternative, suppose that in addition to tabling $\{I_\ell;\ \ell=1,\ldots,m\}$ we elect to table $\{q_{I_1}, q_{I_2},\ldots,q_{I_m}\}$ instead of $\{q_1, q_2\ldots,q_n\}$. To sample X, one proceeds, as before, to select I_L but now calculates $q_{I_{L+1}}$, $q_{I_{L+2}}$,... as needed. Such calculations may be faster and more accurate due to the nearby starting value q_{I_L}. The mean cost in time is proportional to

$$\mu_2 = (1 + c_2)EJ_m - c_2 \qquad (3.2)$$

where c_2 is the (computer dependent) relative increase in time required to evaluate q_i for $i \neq I_L$ as compared to the time required for a table lookup of q_{I_L}.

Clearly $\mu_2 \leq \mu_1$ if and only if

$$c_2\, E(J_m - 1) - c_1\, E(J_m | X > n^*)\, pr(X > n^*) \leq 0\ . \qquad (3.3)$$

In addition, the second procedure requires 2m storage locations while the first requires $m+n^*$. For most applications $c_1 > c_2 > 0$ and as n^* increases the left hand side of expression (3.3) monotonically increases from a negative value of $(c_2 - c_1)EJ_m - c_2$,

when $n^* = 0$, to a positive limit of $c_2 EJ_m - c_2$, if $EX < \infty$. Again, the user is faced with a tradeoff between time and space requirements.

These remarks are intended to be suggestive rather than definitive. Individual decisions should be guided by the requirements and resources of the application intended. Although the dominance of any of the methods described here with respect to time and storage remains a question, one should keep in mind that all the cutpoint methods described preserve the monotone property.

4. ALGORITHMS

In this section we present algorithms for implementing the cutpoint method when sampling from the d.f. of the random variable X. It is assumed throughout that Q is the name of an array or function such that $Q(i) = q_i$. In addition $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

Given the positive integer $M=m$, the algorithm CMSET in Figure 1 returns the array $\{I(\ell) = I_\ell;\ \ell=1,\ldots,M\}$, as described in Section 2. If desired, the array $\{QI(\ell) = q_{I_\ell};\ \ell=1,\ldots,M\}$ suitable for use as described in Section 3 is also returned; otherwise, one deletes statement 8 of CMSET. Note that if the d.f. of X is tabled, the array Q is not destroyed by CMSET.

Figure 1 About Here

Algorithm CM in Figure 2 enables one to sample for X, once the setup in algorithm CMSET has been effected. An algorithm for sampling from the uniform distribution on [0,1) is used in the

first step of CM. The random variable X need not have finite support in order for CM to function correctly. The value of X upon return from CM is the variate desired.

Figure 2 About Here

Algorithms for the setup and use of the alias method are given in Kronmal and Peterson (1979) and Ahrens and Kohrt (1981). Care must be taken with these algorithms to preserve $\{q_i\}$ and avoid the use of large arrays during initialization.

TABLE 1. EXAMPLE OF CUTPOINT METHOD

i	p_i	q_i	m	m+n	$C_{m,n}$	EJ_m
1	.01	.01	1	9	8.00	5.31
2	.04	.05	2	10	4.50	3.31
3	.07	.12	3	11	3.33	2.31
4	.15	.27	4	12	2.75	2.06
5	.28	.55	5	13	2.40	1.71
6	.19	.74	6	14	2.17	1.64
7	.21	.95	7	15	2.00	1.45
8	.05	1.00	8	10	1.88	1.44
			16	24	1.44	1.19

FIGURE 1. ALGORITHM CMSET

```
1.  L ← 0.
2.  J ← 0.
3.  J ← J+1.
4.  QJ ← M*Q(J).
5.  IF QJ ≤ L THEN GO TO 3.
6.  L ← L+1.
7.  I(L) ← J.
8.  QI(L) ← QJ/M.
9.  IF L<M THEN GO TO 5.
10. RETURN.
```

FIGURE 2. ALGORITHM CM

```
1.  SAMPLE U FROM UNIFORM [0,1).
2.  X ← I(⌈M*U⌉).
3.  IF U ≤ Q(X) THEN RETURN.
4.  X ← X+1.
5.  GO TO 3.
```

5. References

Ahrens, Joachim H. and Ulrich Dieter (1973), "Non-Uniform Random Numbers," unpublished manuscript, Institut fur Mathematische Statistik, Technische Hochschule in Graz, Austria.

Ahrens, J. H. and K. D. Kohrt (1981), "Computer Methods for Efficient Sampling from Largely Arbitrary Statistical Distributions," Computing, 26, 19-31.

Chen, Hui-Chuan, and Yoshinori Asau (1974), "On Generating Random Variates from an Empirical Distribution," AIIE Transactions, 6, 163-166.

Fishman, George S. and Baosheng D. Huang (1980), "Antithetic Variates Revisited," Technical Report 80-4, Curriculum in Operations Research and Systems Analysis, University of North Carolina at Chapel Hill.

Fishman, George S. (1978), Principles of Discrete Event Simulation, John Wiley and Sons.

Hoeffding, Wassily (1940), "Masstabinvariante Korrelationstheorie," Schriften des Mathematischen Instituts and des Instituts fur Angewandte Mathematik der Universitat Berlin, 5, 197-233.

Kronmal, Richard A. and Arthur V. Peterson, Jr. (1979), "On the Alias Method for Generating Random Variables from a Discrete Distribution," The American Statistician, 4, 214-218.

Walker, Alastair J. (1974a), "Fast Generation of Uniformly Distributed Pseudorandom Numbers with Floating Point Representation," Electronic Letters, 10, 553-554.

_____ (1974b), "New Fast Method for Generating Discrete Random Numbers with Arbitrary Frequency Distributions," Electronic Letters, 10, 127-128.

_____ (1977), "An Efficient Method for Generating Discrete Random Variables with General Distributions," ACM Transactions on Mathematical Software, 3, 253-256.

Whitt, Ward (1976), "Bivariate Distributions with Given Marginals," Ann. Stat., 4, 1280-1289.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 81-7	2. GOVT ACCESSION NO. AD-A111 967	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Sampling From A Discrete Distribution While Preserving Monotonicity		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) George S. Fishman and Louis R. Moore, III		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0302
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of North Carolina Chapel Hill, North Carolina 27514		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Operations Research Program Office of Naval Research Arlington, VA		12. REPORT DATE February, 1982 13. NUMBER OF PAGES 13
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

sampling
alias method
cutpoint method

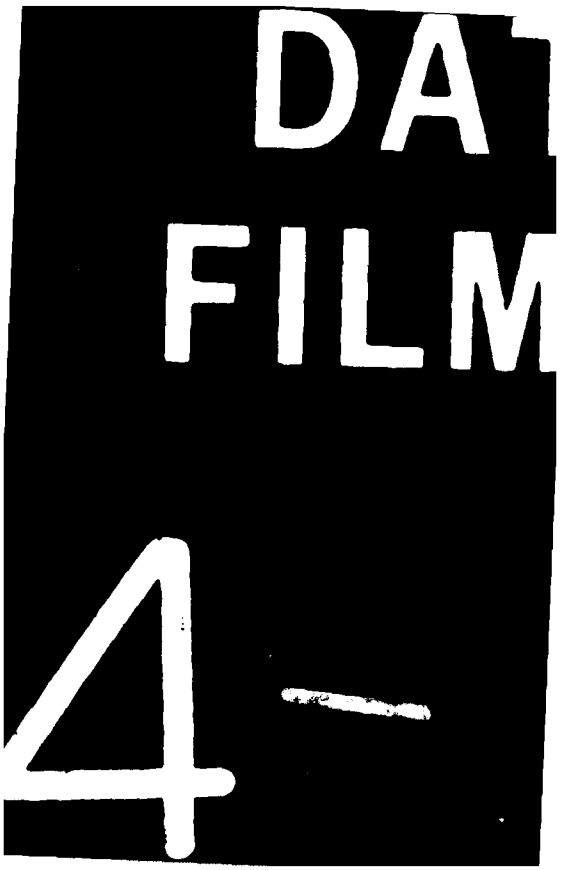
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper describes a cutpoint method for sampling from an n-point discrete distribution that preserves the monotone relationship between a uniform deviate and the random variate it generates. This property is useful when developing a sampling plan to reduce variance in a Monte Carlo or simulation study. The alias sampling method generally lacks this property and requires 2n storage locations while the proposed cutpoint sampling method requires m+n storage locations, where m denotes the number of cutpoints. The

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. expected number of comparisons with this method is derived and shown to be bounded above by $(m + n - 1)/n$. The paper describes an algorithm to implement the proposed method as well as two modifications for cases in which n is large and possibly infinite.


DA
FILM
4