The Annals of Mathematical Statistics
1972, Vol. 43, No. 3, 860-864

UNBIASED COIN TOSSING WITH DISCRETE RANDOM VARIABLES[1]

BY MEYER DWASS

Northwestern University

Hoeffding and Simons (*Ann. Math. Statist.* (1970) 341–352) show how to convert a sequence of i.i.d. indicator random variables with parameter p into such a sequence with parameter $p = \frac{1}{2}$. In the present paper an extension is made to generating sequences of i.i.d. indicator random variables with parameter $p = \frac{1}{2}$ from arbitrary sequences of i.i.d. *discrete* random variables. Also considered is the generating of sequences of i.i.d. r-valued, equiprobable random variables. Some optimality criteria are established.

1. Introduction. In an interesting recent paper, [1], Hoeffding and Simons show various ways of converting a sequence of i.i.d. Bernoulli variables with unknown parameter p, $(0 < p < 1)$, into a sequence of i.i.d. Bernoulli variables with parameter $p = \frac{1}{2}$. Our purpose is to extend some of the calculations in [1] to the case where an arbitrary sequence of nondegenerate, i.i.d. discrete random variables plays the role of the original Bernoulli sequence. In particular, we consider counterparts of the *even* procedures introduced in [1] as well as *type r* procedures for r prime.

2. Paths. We suppose henceforth that $X_1, X_2, \cdots$ is a sequence of nondegenerate i.i.d. random variables with possible values $a_1, a_2, \cdots$. For our purposes it will involve no loss of generality to suppose that the possible values are $1, 2, \cdots$. There may be finitely many or infinitely many possible values. We are motivated by [1] to define a type 2 procedure as follows;

DEFINITION. A 2-valued procedure (N, Z) is a stopping time N and an indicator random variable

$$Z = Z(X_1, \cdots, X_N)$$

such that

$$P(Z = 0) = P(Z = 1) = \tfrac{1}{2}$$

An r-valued procedure is similarly defined, except that the random variable Z assumes r distinct values with equal probabilities. Until Section 6 we will consider 2-valued procedures only, and we refer to them simply as *procedures*.

Suppose that n_j counts the number of times that the value j appears among the n values $X_1, \cdots, X_n$. Then $n_1 + n_2 + \cdots = n$, there being only a finite

Received February 19, 1971; revised October 15, 1971.

[1] Research partially supported by National Science Foundation.

Key words and phrases: Generating unbiased coin tosses, generating random numbers unbiased coin tosses random numbers.

AMS subject classification numbers: 6090.

number of nonzero n_i's in this sum. Denote

$$\frac{n!}{n_1!\, n_2! \cdots} = c(n; n_1, n_2, \cdots)$$

$$c(0; 0, 0, \cdots) = 1 .$$

The multinomial coefficient $c(n; n_1, n_2, \cdots)$ counts the number of distinct assignments of values of n_1 1's, n_2 2's, $\cdots$ to the n random variables $X_1, \cdots, X_n$. Any such assignment is called a *path* with *end-point* $(n; n_1, n_2, \cdots)$. For a given procedure (N, Z) and a given path end-point $(n; n_1, n_2, \cdots)$, it may be that some of the $c(n; n_1, n_2, \cdots)$ paths will lead to the process being stopped just at that point, that is, $N = n$. Let E denote the set of end-points $(n; n_1, n_2, \cdots)$ which are the end-points of at least one path which has just been stopped at the nth step. Let $s(n; n_1, n_2, \cdots)$ be the number of such just-stopped paths with the indicated end-point. Define $s(n; n_1, n_2, \cdots) = 0$ if the end-point is not in E. From now on, we consider only procedures satisfying the following assumption:

ASSUMPTION. It is assumed henceforth that any path with end-point in E has either just been stopped at time n or has already been stopped at an earlier time point. In other words, *all* previously unstopped paths which reach a point in E have to stop at time n.

3. Even procedures. Following the definition for Bernoulli random variables in [1] we define even procedures as follows:

DEFINITION. The procedure (N, Z) is said to be *even* if

(a) $s(n; n_1, n_2, \cdots)$ is always an even integer and

(b) Z equals 1 for half of these $s(n; n_1, n_2, \cdots)$ just-stopped paths and Z equals 0 for the remaining half.

The following characterizes even procedures:

LEMMA 3.1. *The stopping time N permits an even procedure (N, Z) if and only if $c(n; n_1, n_2, \cdots)$ is an even integer whenever the path end-point $(n; n_1, n_2, \cdots)$ is in E.*

PROOF. The following relation holds whenever $(n; n_1, n_2, \cdots)$ is in E.

$$(1) \quad c(n; n_1, n_2, \cdots) = \sum s(m; m_1, m_2, \cdots) c(n - m; n_1 - m_1, n_2 - m_2, \cdots)$$

where the summation is over all end-points $(m; m_1, m_2, \cdots)$ satisfying

$$1 \leqq m \leqq n;\ 0 \leqq m_1 \leqq n_1,\ 0 \leqq m_2 \leqq n_2, \cdots .$$

Since $s(m; m_1, m_2, \cdots)$ is always even, if we assume the procedure is even, it follows that $c(n; n_1, n_2, \cdots)$ is even. The converse can be proved by an induction in n based on writing the right side of (1) as

$$\sum_{1 \leqq m \leqq n-1} + s(n; n_1, n_2, \cdots) .$$

If $(n; n_1, n_2, \cdots)$ is in E, then by the induction hypothesis each of the terms in $\sum_{1 \leqq m \leqq n-1}$ is even. Remember that for the converse we are assuming that

$c(n; n_1, n_2, \cdots)$ is even for end-points in E. Hence, we conclude that $s(n; n_1, n_2, \cdots)$ is even. It remains to check that the assertion holds for some initial value of n. Let n' be the smallest value of n for which the end-point is in E. For any such end-point, we have that

$$s(n'; n_1, n_2, \cdots) = c(n'; n_1, n_2, \cdots)$$

which completes the proof.

4. The best even procedure. Consider a procedure (N', Z') defined as follows: Let $(n; n_1, n_2, \cdots)$ be the end-point of the path determined by the n random variables $X_1, \cdots, X_n$. Let N' be the first index n such that $c(n; n_1, n_2, \cdots)$ is a positive even integer. Clearly N' is a stopping time. Since $c(n; n_1, n_2, \cdots)$ is even for every stopped path, it follows from Lemma 3.1 that $s(n; n_1, n_2, \cdots)$ is even. The random variable Z' can be arbitrarily defined to be 1 for any half of these $s(n; n_1, n_2, \cdots)$ paths. Notice that the assumption at the end of Section 2 is satisfied.

The procedure (N', Z') is *best* among even procedures in the sense that if (N, Z) is any other even procedure, it follows immediately from the definition of N' and from Lemma 3.1, that with probability 1, $N' \leqq N$.

5. The generating function for N'. In [1] it was shown that when the X_i's are Bernoulli random variables with parameter p, then

$$\sum_{n=0}^{\infty} P(N' > n)t^n = \prod_{k=0}^{\infty} [1 + (pt)^{2^k} + (qt)^{2^k}]. \tag{2}$$

An extension to arbitrary discrete X_i's will be proved along very similar lines. The counterpart of (2) is the following:

THEOREM 5.1. *Suppose that* $P(X_1 = i) = p_i$, $i = 1, 2, \cdots$. *Then*

$$\sum_{n=0}^{\infty} P(N' > n)t^n = \prod_{k=0}^{\infty} [1 + (\sum_{i=1}^{\infty} p_i^{2^k})t^{2^k}].$$

The proof is based on the following lemma about multinomial coefficients:

LEMMA 5.2. *Suppose that* m *is of the form* 2^k, $k = 1, 2, \cdots$. *Then* $c(m; m_1, m_2, \cdots)$ *is even whenever there are at least two nonzero* m_i*'s among* $m_1, m_2, \cdots$.

PROOF. The proof is by induction on k. The result holds for $k = 1$, since $c(2; 1, 1, 0, \cdots) = 2$. By considering the coefficient of $t_1^{m_1} t_2^{m_2} \cdots$ in $(t_1 + t_2 + \cdots)^{2^k}(t_1 + t_2 + \cdots)^{2^k}$ we know that

$$\sum c(2^k; i_1, i_2, \cdots)c(2^k; j_1, j_2, \cdots) = c(2^{k+1}; m_1, m_2, \cdots) \tag{3}$$

where the summation is over all nonnegative indices, $i_1, j_1, i_2, j_2, \cdots$ such that $i_1 + j_1 = m_1$, $i_2 + j_2 = m_2, \cdots$. By hypothesis, the terms on the left side of (3) are all even with the exception of those products which equal one; for example, terms of the form $c(2^k; 2^k, 0, \cdots)c(2^k; 0, 2^k, \cdots)$. But for any such term, there is always a "complementary" term. For instance, the term "complementary" to the one just cited above is $c(2^k; 0, 2^k, \cdots)c(2^k; 2^k, 0, \cdots)$. So the contribution from these terms is also even. Hence, $c(2^{k+1}; m_1, m_2, \cdots)$ is even.

PROOF OF THEOREM 5.1. The proof now follows that in [1] exactly. First we have that by Lemma 5.2,

$$P(N' > 2^k) = p_1^{2^k} + p_2^{2^k} + \cdots, \qquad k = 1, 2, \cdots. \tag{4}$$

Next we have that if $0 \leqq n < 2^k$, then

$$P(N' > 2^k + n) = P(N' > 2^k)P(N' > n). \tag{5}$$

(It follows from Lemma 5.2 that $N' > 2^k + n$ implies that $X_1 = \cdots = X_{2^k}$, from which (5) follows directly.) Thus, (4) and (5) allow an explicit evaluation of $P(N' > r)$ for all r. The generating function argument leading to Theorem 5.1 is now completely routine. (See [1].)

REMARKS. It is clear that if we have a k-faced die we are better off in using the full structure of the die rather than dichotomizing the outcomes into two sets of face types and thus generating Bernoulli trials. For example, with an "honest" 6-faced die, the expected value of the best even procedure described above is

$$EN' = \prod_{k=0}^{\infty}[1 + (\tfrac{1}{6})^{2^k-1}] \approx 2.4.$$

When the die is dichotomized, the best even procedure (whose stopping time is denoted N_2 in [1]) has expected value

$$EN_2 = \prod_{k=0}^{\infty}[1 + (\tfrac{1}{2})^{2^k-1}] \approx 3.4.$$

6. Type r procedures. Let r be a positive integer. We define an r-valued procedure to be of type r if $s(n; n_1, n_2, \cdots)$ is always a multiple of r. Thus, the above-discussed even procedures are of type 2. We suppose that the assumption at the end of Section 2 still prevails.

Lemma 3.1 holds for any r. That is,

LEMMA 6.1. *The stopping time N permits a type r procedure (N, Z) if and only if $c(n; n_1, n_2, \cdots)$ is a multiple of r whenever $(n; n_1, n_2, \cdots)$ is in E.*

PROOF. The proof is the same as that of Lemma 3.1.

We can now describe the best type r procedure, (N'', Z''), just as in Section 4, by letting N'' be the first index n such that $c(n; n_1, n_2, \cdots)$ is a multiple of r, where $(n; n_1, n_2, \cdots)$ is the end-point of the path determined by $X_1, \cdots, X_n$. However, an explicit expression similar to Theorem 5.1 seems to be available only when r is a prime number. The relevant background fact which plays the role of Lemma 5.2 is the following.

LEMMA 6.2. *Suppose that r is a prime number and m is of the form r^k, $k = 1, 2, \cdots$. Then $c(m; m_1, m_2, \cdots)$ is a multiple of r whenever there are at least two nonzero m_i's among $m_1, m_2, \cdots$.*

PROOF. It is easy to check that $c(r; m_1, m_2, \cdots)$ must be a multiple of r when there are two or more nonzero m_i's among $m_1, m_2, \cdots$. Thus the assertion holds

for $k = 1$. An induction proof now follows along the same lines as for Lemma 5.2.

REFERENCE

[1] HOEFFDING, W. and SIMONS, G. (1970). Unbiased coin tossing with a biased coin. *Ann. Math. Statist.* **41** 341–352.