The Frequency Interpretation in Probability

Charles Friedman
Math.Dept. RLM 8.100
Mail Code: C1200
University of Texas
Austin, Texas 78712

email: friedman@math.utexas.edu

Abstract:

We formulate and discuss the notion of *generic sequence* (and *random sequence*) associated with a sequence of random variables. These are substantial generalizations of the notion of "collective" introduced by R.von Mises as an attempt to give an operational meaning to various probabilistic ideas. The definition of collectives lacked a precise formulation, and A. Church and others attempted to give a rigorous meaning to the theory by utilizing ideas from the theory of computability. Nevertheless, the theory had major flaws. Even though some important contributions to special cases were made by P. Martin-Löf, the whole circle of ideas has languished due to lack of generality. The theory presented in the present article resolves many of these problems, and provides a coherent framework for the relevant ideas.

In Section 2, we discuss some of the history and background necessary to our discussion. Section 3 develops the fundamental notions of generic and random sequences and provides an operational meaning for probabilistic notions associated with a sequence of random variables. In Section 4, we discuss some of the implications of the ideas developed.

Section 1: Introduction

The difficulties involved in the interpretation of probabilistic notions in statistics (and in many other areas also) is well known, and the attempts to resolve these problems have by now a substantial history. In *The Foundations of Statistics* ([6]), L. J. Savage notes "... as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement since The Tower of Babel." Closely connected to probabilistic foundations and of especial importance in statistics is the concept of *randomness*, concerning which no less a scholar than H. Cramér admits in his *Mathematical Methods of Statistics* ([2])that "It does not seem possible to give a precise definition of what is meant by the word *random*." Of course, one may define a *random sample from X* (where X is a random variable) as a sequence of random variables which are independent and have the same distribution as X. However, if one has a sequence of data values and wonders in what sense it can be considered *random*, the cited definition is not particularly helpful; indeed, it is somewhat unclear what randomness of a particular sequence even means, (although various approaches to this notion are now known.) Our purpose in the present article is to provide a definite meaning for this notion and for the related notion of the frequency interpretation of probability which seems more general than previous attempts at such definitions. In fact, we shall show that there exist sequences of data values that model "all" relevant properties associated with a sequence of random variables. The adjective "all" in the preceeding will be understood as referring to all properties that hold with probability 1 and *may be described by algorithm.* This notion seems quite useful in the theory of probability, but it seems to have been exploited very infrequently.

Throughout the present article, nothing inconsistent with the classical theory of probability will appear; the only novelty will perhaps be the point of view. (It should be remarked that there are notions of randomness that are distinct from the usual probabilistic one, but we will not be at all concerned with these here.)

Before presenting our theory, we discuss some of the history of these ideas.

Section 2: Background

In order to restrict attention to a specific context, suppose we contemplate tossing a coin. What does it mean to say that the probability of obtaining a head is $\frac{1}{2}$? It might be claimed that this reflects our complete lack of knowledge concerning the outcome; however, it could be that the coin is weighted in a way which siginificantly affects the outcome or it might be 2-headed, *etc*, in which case it seems that the cited probabilistic prediction is just wrong. Additionally, one can argue that there is nothing probabilistic at all occurring; the coin will be acted on by various forces, and the laws of mechanics will determine the outcome. There is probably no easy answer to these objections if we only consider a single toss. There are situations involving a single occurrence when a subjective or "personalistic" concept of probability seems useful: for example, with regard to questions such as "What is the probability that a successful brain transplant will be performed on a human in the next year?" But the subjective approach seems less compelling in the coin tossing situation, although one could make some argument for its applicability. In any case, we shall have nothing more to say concerning this approach. The classical *a priori* approach to the coin example would involve selecting mutually exclusive events which are "equally likely", in this case the events that head or tail occur, and assigning them equal probabilities ($\frac{1}{2}$). Evidently, this contributes nothing new to the understanding of the situation, the term "equally likely" being merely a substitute for the arbitrary assignment of probability $\frac{1}{2}$. The axiomatic point of view (espoused by Kolmogorov and others) avoids any attempt to assign a "meaning" to such probabilistic notions. One considers a sample space, $S = \{H, T\}$, and a function P defined on "events", which is in the present situation determined by the values $P(H) = P(T) = \frac{1}{2}$. This is a purely mathematical setup; it can be used to compute various things, but what it all "means" is an unmeaningful or perhaps inappropriate question.

If instead of a single toss we consider a repetition of this action, it seems that a new interpretation of the probabilities discussed is possible. If we adhere to the frequency (or *a posteriori*) interpretation of probability, we attempt to assign meaning to the assertion

$P(H) = \frac{1}{2}$ by some statement such as "In a long sequence of tosses, we expect the relative frequency of heads (*i.e.* # heads/# tosses) to be about $\frac{1}{2}$." This seems meaningful for a few seconds, but then we realize that the word "expect" involves the probabilistic notions that we failed to understand in the first place now appearing in a more complex manifestation, the word "about" is difficult to quantify, and in addition, a whole new set of unpleasant questions are presented. For example, how is the tossing to be performed? Clearly, if one tosses in exactly the same way each time and the surrounding conditions are identical, then the laws of physics imply that the identical result will always occur. On the other hand, if one tosses in such a way that $\frac{1}{3}$ of the outcomes are heads, then that's what happens; what does this all have to do with probability? Furthermore, even after performing a long sequence of tosses, it seems difficult to come to any probabilistic conclusions without refering to concepts equivalent to those whose meaning we wish to understand. For example, we could toss a coin 1000 times with the result being, say, 463 heads. What then? Does this reinforce the belief that $P(H) = \frac{1}{2}$? Now it is certainly true that there are statistical tests of the hypothesis that $P(H) = \frac{1}{2}$, but these make use of a large body of probabilistic theory; in using these, we can test the hypothesis, but without explaining what the hypothesis means!

For the practical working statistician this may be largely irrelevant, but from a theoretical and foundational viewpoint it appears incomplete and unsatisfying. In a final attempt to resolve the problem under discussion, one might say that for the coin tossing situation $P(H) = \frac{1}{2}$ simply means that we believe the sequence of outcomes will be one for which the asymptotic relative frequency of heads is $\frac{1}{2}$, *i.e.*, we use such a sequence as a *model* for the probabilistic concept. The main problem with such an approach is that many sequences with the given asymptotic property fail to satisfy other conditions that would be expected of a sequence of independent tosses of a fair coin; it seems that to be a good model, "all" such conditions should hold; but is this possible? If not, then the idea of a model is somewhat suspect; one might need different models for many different properties.

In the early 1900's Richard von Mises (see *e.g.* [7] or [4]) proposed a remarkable and innovative framework in which to attack these problems. He suggested that the notion of probability $\frac{1}{2}$ (and other probabilities) as well as the notion of randomness could be defined in terms of properties of a representative sequence of outcomes which he called a "collective". In particular, consider the sequence of oucomes of a series of coin tosses which we represent as an infinite sequence, $\{x_n\}, n = 1, 2, \ldots$, of 0's and 1's, the 1's denoting that a head has occurred, the 0's the contrary. (Of course, in an actual experiment, the outcomes form a finite sequence, but we are imagining a sequence of tosses which continues indefinitely.) In order for such a sequence to be termed a "collective" for the "fair" coin tossing experiment, it must satisfy the property that the asymptotic relative frequency of

1's is $\frac{1}{2}$. This is not enough, however, since a sequence such as: 1 0 1 0 1 0 1 0 1 0... satisfies this asymptotic condition, but does not have the quality of "randomness" that one expects. von Mises also required that for any method of selecting a subsequence, x_{n_i} of the x_n (with $n_1 < n_2 < \ldots$) in which the choice to include or not include a given x_n can depend only on x_i with $i < n$, the asymptotic relative frequency of 1's in the chosen subsequence should again be $\frac{1}{2}$. The idea is that a gambler making bets at even odds on whether a given next toss will be heads or not should not be able to formulate any scheme involving when to bet which would provide an advantage. (Such a scheme might be something of the sort: "whenever 5 heads occur consecutively bet that the next is not a head, and if head and tail alternate for more than 10 tosses, don't bet for the next 5 tosses", *etc.*) It would be very nice if this made sense, because then one could present the collective as a *model* of the concept "a random sequence of 0's and 1's with probability $\frac{1}{2}$ of occurrence of 1." The problem is that if "all" schemes are allowed, then there are no collectives because there is always a choice of subsequence for which the asymptotic relative frequency of 1's differs from $\frac{1}{2}$ - just pick the terms equal to 1! The notion that the choice to include or not include a given x_n "can depend only on x_i with $i < n$" really has no meaning in this situation; what does "depends only on" mean? Think of the situation of a gambler deciding whether and how to bet on the next toss; he might get lucky and always guess to his advantage. An attempt at a solution to this difficulty was suggested by Alonzo Church ([1]): the choice of subsequence should be required to be given by an *algorithm* (an effectively computable function, *i.e.*, one which can be coded in some programming language.) In this case, the notion"depends on" makes good sense; a computed result depends only on some numbers if it uses them (and nothing else) as input. It follows rather easily from the fact that the totality of algorithms is *countable* that, with this definition, collectives exist (and form a set of probability 1 with respect to the canonical probability measure in this situation.) However, it was soon pointed out by J. Ville (see [4] for some discussion and references) that collectives in the above sense exist for which the frequency of 1's in the first n terms is always $\geq \frac{n}{2}$; unhappily, this is a property which holds with probability 0 (again with respect to the canonical probability.) Similarly, there are collectives for which the Law of the Iterated Logarithm fails to hold. These objections essentially put an end to the attempts to offer collectives as a model of probabilistic notions. In retrospect, it seems clear that the definitions made were quite inadequate in that they focused on a few properties that ought to hold for random sequences but omitted many others. For example, the restriction to subsequences x_{n_i} with $n_1 < n_2 < \ldots$ is artificial in general (although perhaps appropriate in the situation envisioned of a gambler betting on a sequence of coin tosses); there is no reason to exclude "subcollections" x_{n_i} generated by an algorithm for which the n_i are merely assumed to be distinct and for which the successive choices of terms depend only on previously examined terms. In addition, it ought to be the case that if such a subcollection is chosen (algorithmically), then not only should the asymptotic relative frequency of 1's still be $\frac{1}{2}$, but if one looks at successive blocks of terms of some fixed length k, say, these should be asymptotically uniformly distributed over the 2^k available possibilities. Even if one required such conditions to hold, it would additionally be necessary to exclude examples of the type exhibited by Ville; these exhibit "unlikely" behavior which is not excluded by conditions on asymptotic *relative* frequencies. Furthermore, none of these ideas apply directly to more general situations in which one has a sequence of data points which are *real* (rather than just 0's and 1's); we would also like to make sense of the notion of "generic" sequences which have all "expected" properties of the sampled values of general sequences of random variables (which might not be independent, for example.) We will, in fact, address this general situation in the sections to follow. For the case of binary sequences (and somewhat more generally), however, there is a theory which provides a meaning for some of the probabilistic notions discussed; this is the theory of Kolmogorov (or algorithmic) complexity (see [4] for an extremely extensive coverage.) We briefly describe this idea as discussed in an article of P. Martin-Löf ([5]).

Let x be a sequence of 0's and 1's, and let A be a partial recursive function (a "program".) The complexity of x relative to A is the minimum number of dyadic symbols necessary to compute x using A; symbolically, $K_A(x) = \min_{A(p)=x} l(p)$. This depends on A, but if one chooses A to be a universal partial recursive function (the partial recursive function computed by a universal Turing machine), then for any other partial recursive B, $K_A(x) \leq K_B(x) + C$, where C is a constant depending only on A and B. Let A be such a universal function and denote K_A by K. One then defines a sequence $\{x_k\}, k = 1, 2, \ldots$ of 0's and 1's to be "random" if $K(x_1, \ldots, x_n) \geq n - C$ for infinitely many n. (One might contemplate requiring this for all n, but it turns out that no sequences would satisfy such a condition. There are alternate definitions of complexity with respect to which such a condition is appropriate. See [4] for a complete theory.) Such a definition is appealing in many ways; the random sequences are defined to be the ones which are "most" difficult to describe algorithimically. Martin-Löf states the theorem that the random sequences in the sense just defined belong to the complement of all *constructively measurable* sets of measure 0. (A subset N of the binary sequences is a constructively measurable set of measure 0 if for each $\epsilon > 0$ thereis a recursive function f on the natural numbers whose values are finite binary sequences and $N \subseteq U_\epsilon = \cup_{k=1}^{\infty} U(f(k))$ where $U(a_1, \ldots, a_n)$ is the set of sequences beginning with $a_1, \ldots, a_n$ and U_ϵ has (product) measure $< \epsilon$.) Now an *effective statistical test of randomness* can be identified with a property which holds for all infinite binary sequences except those in a constructively measurable set of measure 0 (canonical measure still), so it follows that the random sequences obey all effective (*i.e.*, given by algorithm) statistical tests of randomness. However, as mentioned above, there is some difficulty in extending both notions- the K complexity and Martin-Löf's idea of requiring all effective statistical tests of randomness to hold- to sequences of values of more general random variables. We would like to be able to define a notion of "generic" sequence of values associated with a sequence of random variables which would serve as a model for all probabilistic/statistical notions associated with the given random variables. We shall do this in the next section in a manner inspired by both von Mises and Martin-Löf. It is interesting to note that H. Cramér in his classic text *Mathematical Methods of Statistics* ([2]) comes close to enunciating the notion we have in mind, at least in the case of sequences of independent random variables (although in the final analysis, one must probabably consider his point of view to be the axiomatic one mentioned earlier.) We will quote from Cramér's text, changing notation slightly in a few places (*e.g.*, replacing German script letters by Latin counterparts.) Cramér enunciates the frequency interpretation of probability in no uncertain terms, in stating "Whenever we say that the probability of an event E with respect to an experiment $\mathcal{E}$ is equal to P, the concrete meaning of this assertion will thus simply be the following: In a long sequence of repetitions of $\mathcal{E}$, it is practically certain that the frequency of E will be approximately equal to P." A bit later he considers the notion of random variable: "Consider a determined random experiment $\mathcal{E}$, which may be repeated a large number of times under uniform conditions. We shall suppose that the result of each particular experiment is given by a certain number of real quantities $\xi_1, \xi_2, \ldots, \xi_k$, where $k \geq 1$. We then introduce a corresponding variable point or vector $\xi = (\xi_1, \xi_2, \ldots, \xi_k)$... We shall call ξ a k-dimensional *random variable*. Each performance of the experiment $\mathcal{E}$ yields as its result an *observed value* of the variable ξ..." There appears to be almost an *identification* of the "random variable" with the sequence of values obtained by experiment (rather than with some *function* on a sample space as in the axiomatic approach.) Next we are told "Let S denote some set...and let us consider the event $\xi \subseteq S$. We shall assume that this event has a definite probability P in the sense explained in 13.5" (the frequency interpretation described above.) "The number P will...be denoted by... $P = P(S) = P(\xi \in S)$." A paragraph later an axiom is stated to the effect that "To any random variable ξ...there corresponds a set function $P(S)$ uniquely defined for all Borel sets S...such that $P(S)$ represents the probability...of $\xi \in S$." It is not exactly clear how to interpret this, but if taken at face value it seems difficult to escape the conclusion that it is being asserted that all probabilities of the form $P(\xi \in S)$ can be obtained from the *empirical distribution* of the sequence of data values, *i.e.*, as asymptotic relative frequencies associated with the sequence. This, of course, is too much to expect. It is true that if we have a sequence of independent, identically distributed, random variables (or vectors) X_i, then for each measurable set S, almost all sample sequences have the property that the asymptotic relative frequency with which its terms belong to S is $P(X_i \in S)$. (This follows from the strong law of large numbers as will be discussed later.) However, there are uncountably many S, and this likely precludes the existence of sequences for which the condition holds for *all* (Borel) S. In fact, if we take for S the set of points in the given sequence of data values, then necessarily the asymptotic relative frequency with which these data points belong to S will be 1! However, in general, it won't be the case that this (countable) set S has probability 1. We could eliminate this difficulty by dealing with only countably many S, but as indicated earlier, this still won't provide a theory in which data sequences exist having all desirable properties, many of which depend on properties more delicate than asymptotic relative frequencies(recall the discussion of Ville's examples above.)

In the next section we will show that associated with a sequence of random variables there exist sequence of data values which have "all" expected generic properties, and in fact such sequences have probability 1.

Section 3: Generic Sequences

In the present section we shall use the term *algorithm* in its usual sense - *i.e.* referring to an algorithm representable as a (finite) program in some programming language in which general recursive functions may be computed. (This includes all standard languages.) We often use the term *program* as a synonym for *algorithm.* The programs of interest here will be ones which take real numbers as input. Of course, using standard encoding and storage of numbers, only a finite precision is possible so, in fact, all input will consist of *rational* numbers. We do not place any bounds on the precision used, but this is always assumed to be either built into the program or determined by the form of the the numbers input. For example, it might be determined that all input is rounded in some specific way (*i.e.* up or down) to 10 digits by the program; alternately, it could be that the input is rounded to 10 digits before it is accessed by the program. In any case, the precision is assumed to be part of the algorithm or program. This is just the situation which must hold in practice for any computations performed on real computing devices. Whenever we speak about the input of some real number to a program, we mean that the number is accessed by the program according to the precision implicit in the program. However, (and this is crucial), we don't restrict the real numbers under discussion to be *computable*, *i.e.*, we do not require that there is an algorithm which computes the numbers to arbitrary precision. (If we consider some number x which is not computable, it still makes sense to contemplate the behavior of programs when input various approximations of x, even if such approximations are not actually obtainable by an algorithm.)

Now suppose that we have a sequence of random variables $X_1, X_2, X_3, \ldots$ on some probability space $\{S, P\}$ (*i.e.* S is a sample space, and P is a probability on S.) Suppose also that $x_1, x_2, x_3, \ldots$ is a sample sequence (*i.e.* $x_i = X_i(s)$, some $s \in S$.) We want to define what it means for $x_1, x_2, x_3, \ldots$ to be a *generic* sequence for the sequence of random variables $X_1, X_2, X_3, \ldots$. Consider a program which behaves in the following way: Suppose $a_1, a_2, \ldots$ is a sequence of real numbers (some of) the terms of which will be used as inputs to the program. The program first outputs a natural number n_1 indicating that a_{n_1} should be input after which it computes and outputs a natural number n_2 indicating that a_{n_2} is to be input next, then computes a natural number n_3 indicating that a_{n_3} should be input, *etc.* We don't assume that the n_j are increasing or even distinct (although the case of distinct n_j is probably the most interesting and relevant one.) Assume also that sometimes after a_{n_j} has been input, the program computes and outputs some value $f_j(a_{n_1}, \ldots, a_{n_j})$, where f_j is a function on some domain in $\mathbb{R}^j$. Suppose that it is the case that

$$P(\limsup_{j\to\infty} f_j(X_{n_1}, \ldots, X_{n_j}) = \alpha) = 1 \tag{3.1}$$

where α is a real number. (Note that the indices j form an increasing subsequence of the natural numbers, and the limit as $j \to \infty$ is taken as j increases through this subsequence. It is not assumed that infinitely many f_j are computed for all sequences a_i, but this is, of course, assumed to be true for almost all sample sequences. Note also that the f_j computed and the n_j depend on the sample sequence.)

Definition 3.2

The sequence $x_1, x_2, x_3, \ldots$ is *generic* for the sequence of random variables $X_1, X_2, X_3, \ldots$ if for all algorithms of the type described and satisfying a condition of the form (3.1), we have

$$\limsup_{j\to\infty} f_j(x_{n_1}, \ldots, x_{n_j}) = \alpha. \tag{3.3}$$

From (3.1) and the fact that there are only countably many algorithms it follows that *the generic sequences have probability* 1 (or more properly, the sample points $s \in S$ for which $X_1(s), X_2(s), \ldots$ is a generic sequence have probability 1.)

The generic sequences are precisely those that satisfy all conditions expressed by algorithm that are satisfied with probability 1 in the set of sample sequences for the sequence of random variables $X_1, X_2, X_3, \ldots$. This statement is one that undoubtedly can't be "proved" since the meaning of the phrase "all conditions expressed by algorithm" is not definitely specified, but a little thought indicates that the claim is justified.

Remark

The definition of generic sequence can be phrased in many different but equivalent ways. The precise form of the equations (3.1), (3.2) has been chosen for convenience and ease of application. Obviously, one could replace the "lim sup" with "lim inf" in those equations with no change in meaning since $\limsup_{n\to\infty} a_n = -\liminf_{n\to\infty}(-a_n)$. It follows that if $x_1, x_2, \ldots$ is a generic sequence for the sequence of random variables $X_1, X_2, \ldots$, and we have a program of the type described above for which

$$P(\lim_{j\to\infty} f_j(X_{n_1}, \ldots, X_{n_j}) = \alpha) = 1$$

then we have

$$\lim_{j\to\infty} f_j(x_{n_1}, \ldots, x_{n_j}) = \alpha.$$

In order to relate the notion of generic sequence with those considered in the previous section, we consider the particular case of generic sequences associated with a sequence of independent, identically distributed random variables.

Definition 3.4

Suppose $X_1, X_2, \ldots$ are independent, and identically distributed with common distribution dF (*i.e.* $P(X_i \le x) = F(x)$) which we (will always) assume has finite expectation and variance. In this case a sequence $x_1, x_2, \ldots$ which is generic for the X_i is called a *random sequence* associated with dF.

It will often be the case that dF is a uniform distribution (either on some finite subset or an interval of the real numbers; in the former case dF is usually termed a *discrete uniform distribution*; such a dF assigns equal weight to each point in the subset.)

Now consider the case of a random sequence $x_1, x_2, \ldots$ associated with the uniform distribution on the two point set $\{0, 1\}$. We claim that such a sequence provides the correct generalization of the notion of "collective" introduced by von Mises. First, it is clear that the asymptotic relative frequency of 0's (or 1's) in the sequence of x_i is $\frac{1}{2}$ since one could take $f_j(x_1, \ldots, x_j) = \frac{1}{j}(x_1 + \ldots + x_j)$ in (3.1), (3.3). (There is a subtlety concerning the storage space available for such a program which we shall discuss below.) However, von Mises required that the asymptotic relative frequency of 0's should be $\frac{1}{2}$ for any subsequence x_{n_i} selected in such a way that the choice to include or exclude a given x_n depends only on x_i with $i < n$. We remarked earlier that this notion of "depends only on x_i with $i < n$" makes sense if one requires that the choice of subsequence is defined by an algorithm. Now we may easily contemplate an program of the type described above in connection with the definition of generic sequences which takes as input the terms of a sequence a_i of 0's and 1's and selects a subsequence a_{n_i} with $n_1 < n_2 < \ldots$ where each n_i is computed using as input only terms a_i with $i < n_i$. In addition, the program computes the expression $f_j = \frac{1}{j}(a_{n_1} + \ldots a_{n_j})$ after each a_{n_j} has been input. Now we want it to be the case for generic sequences x_i that

$$\lim_{j\to\infty} \frac{1}{j}(x_{n_1} + \dots x_{n_j}) = \frac{1}{2} \tag{3.5}$$

(where the n_j are the indices computed by the program using the x_j as input.)
For this to hold we would need that

$$P(\lim_{j\to\infty} \frac{1}{j}(X_{n_1} + \dots X_{n_j}) = \frac{1}{2}) = 1. \tag{3.6}$$

In this last equation the n_j depend on the sample point $s \in S$, *i.e.* are *random variables.* Why does (3.6) hold? The relevant fact is the following result.

Theorem 3.7

Suppose $X_1, X_2, \dots$ are independent and identically distributed with common distribution dF (*i.e.* $P(X_i \le x) = F(x)$) as in **3.4**. Let $\mathcal{F}$ be the sigma algebra of measurable sets in the underlying sample space S for the process, and let $\mathcal{F}_k = \mathcal{F}(X_1, X_2, \dots, X_k)$ be the sub sigma algebra generated by $X_1, \dots, X_k$ (*i.e.*, $\mathcal{F}_k$ is the smallest sub sigma algebra of $\mathcal{F}$ with respect to which $X_1, \dots, X_k$ are measurable.) Suppose that $n_1, n_2, \dots$ are natural number valued random variables satisfying $n_1 < n_2 < \dots$ pointwise with probability 1, $\{n_i = k+1\} \in \mathcal{F}_k$ for each i and $k \ge 1$, and $\{n_1 = 1\}$ is independent of $\mathcal{F}_1$. Then the $X_{n_i}, i = 1, 2, \dots$ are independent and identically distributed with the same distribution as the X_i.

Proof(sketch):

$$P(X_{n_i} \in A_i, i = 1, \dots, m) = \sum_k P(X_{n_i} \in A_i, i = 1, \dots, m,\ n_m = k) =$$
$$\sum_k P(X_{n_i} \in A_i, i = 1, \dots, m-1,\ X_k \in A_m,\ n_m = k) =$$
$$\sum_k P(X_{n_i} \in A_i, i = 1, \dots, m-1,\ n_m = k) \cdot P(X_k \in A_m) =$$
$$P(X_{n_i} \in A_i, i = 1, \dots, m-1) \cdot P(X_k \in A_m)$$

where the third equality follows from the fact that the set $\{X_k \in A_m\}$ is independent of $\{n_m = k\}$. (Note that $P(X_k \in A_m)$ is a constant independent of k.) Hence, the result claimed follows by induction on m.

We remark that the article of Church ([1]) mentioned previously makes use of a theorem of Doob ([3]) which is very similar to the one just proved.

Remark

In the previous discussion we have considered a program which computes functions $f_j(x_1,\ldots,x_j)=\frac{1}{j}(x_1+\ldots+x_j)$, *etc.* In the situation being considered the x_i are 0 or 1, and the value of f_j is a rational number. Now it is generally assumed that the computing machines on which our programs are executed have unlimited storage and workspace so that arbitrarily large inputs can be accepted and operated on. The circumstance that, in fact, there is no such machine with unbounded storage does not really create any logical difficulty since one can *imagine* the result of a computation on such a machine without actually performing it. However, we don't assume that the values of the functions f_j are computed with complete precision (although in the case of rational numbers this could be the case if unbounded storage and workspace is assumed.) Rather, it may be that rationals are converted to reals and computed in some bounded precision as discussed earlier. If it is the case that $f_j(x_1,\ldots,x_j)$ is only an approximate computation of $\frac{1}{j}(x_1+\ldots+x_j)$ (still considering the case of a random sequence $x_1,x_2,\ldots$ associated with the uniform distribution on the two point set $\{0,1\}$), then it will not necessarily be the case that $P(\lim_{j\to\infty} f_j(X_1,\ldots,X_j)=\frac{1}{2})=1$. However, this problem disappears if we consider instead a program which computes values $f_j(x_1,\ldots,x_j)$ which are equal to $\frac{1}{2}$ whenever the finite precision value obtained for $\frac{1}{j}(x_1+\ldots+x_j)$ is within some fixed distance ϵ of $\frac{1}{2}$ and 0 otherwise. (ϵ could be any small rational whose precision does not exceed that allowed and which is greater than the error produced in rounding the values computed for the $\frac{1}{j}(x_1+\ldots+x_j)$.) What we are saying here is that since approximate calculation of the asymptotic relative frequency of 1's will yield a value approximately equal to $\frac{1}{2}$ for almost all sample sequences (strong law of large numbers), this property will hold for particular random (*i.e.* generic) sequences also. Since this must hold for any degree of approximation, the asymptotic relative frequency of 1's in random sequences must be $\frac{1}{2}$. The point is that there is no logical difficulty in the definition of generic sequence; one must just consider appropriate algorithms when considering the consequences of the definition.

A situation and reasoning similar to the above is relevant to the case of more general independent, identically distributed processes X_i (whose values may be real numbers which are utilized by our algorithms only in some finite precision) when considering consequences of the strong law of large numbers for values such as $\lim_{j\to\infty}\frac{1}{j}(x_1+\ldots+x_j)$, *etc*, where $x_1,x_2,\ldots$ is a generic sequence. In such a situation only approximations of the values $\frac{1}{j}(x_1+\ldots+x_j)$ can be computed by our programs, but the resolution of this apparent difficulty is achieved as in the case discussed above.

We will, in what follows, sometimes omit detailed discussion of the nature of the appropriate "approximate" algorithms which must be considered in conjunction with the definition of generic sequence in order to deduce some property of such sequences, especially when the reasoning seems routine and obvious.

We now discuss some further properties of random sequences $x_1, x_2, \ldots$ associated with the uniform distribution on the two point set $\{0, 1\}$. Some of these concern adjacent blocks of length j: $x_{(k-1)j+1}, \ldots, x_{kj}$, ($j$ fixed, $k = 1, 2, \ldots$). We could just as well consider blocks of length j from a subsequence x_{n_i} where the choice to include or exclude a given x_n depends only on x_i with $i < n$ by invoking **Theorem 3.7**, but for the sake of simpler notation we shall not do this; it should, however, be kept in mind that such a generalization holds for the examples to be discussed now and for various later examples.

First, it follows from our definition of *random sequence* and the Strong Law of Large Numbers that if we define the sequence of vectors $\mathbf{w}_k = (x_{(k-1)j+1}, \ldots, x_{kj})$, $k = 1, 2, \ldots$, then the *empirical distribution* of the $\mathbf{w}_k$ is uniform on the set of 2^j binary (*i.e.* with entries 0 or 1) vectors of length j. That is, if $\mathbf{e}$ denotes any of these binary vectors,

$$\lim_{N\to\infty} \frac{1}{N} \cdot \#\{\mathbf{w}_k, k = 1, \ldots, N | \mathbf{w}_k = \mathbf{e}\} = 2^{-j}. \tag{3.8}$$

(Apply **Definition 3.2** with $f_{Nj}(x_1, \ldots, x_{Nj}) = \frac{1}{N} \cdot \#\{\mathbf{w}_k, k = 1, \ldots, N | \mathbf{w}_k = \mathbf{e}\}$.)
It is not clear to the present author whether or not the "collectives" in the sense of von Mises and Church discussed above must satisfy (3.8). (In any case, we will see below that other properties that should hold for "random" sequences and which do not generally hold for collectives do hold for random sequences as we have defined them.)

Next we indicate in what sense random sequences satisfy the Central Limit Theorem. In order to do this, we first must phrase this theorem in a form involving limits of the type in (3.1), which is perhaps a bit unexpected since the usual formulation of the Central Limit Theorem involves convergence in distribution. Suppose $X_1, X_2, \ldots$ are independent and identically distributed with expectation μ, variance σ^2 and distribution $dF = P_X$. For each $j \geq 1$, $k \geq 1$, let

$$W_{j,k} = \frac{1}{\sqrt{j}\sigma} \sum_{i=(k-1)j+1}^{kj} (X_i - \mu). \tag{3.9}$$

Let A be a real interval, and denote the characteristic (indicator) function of A by 1_A. For fixed j, the $1_A(W_{j,k})$ are independent and identically distributed, and by the Strong Law of Large Numbers

$$\lim_{N\to\infty} \frac{1}{N} \#\{W_{j,k} \in A,\ k \leq N\} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} 1_A(W_{j,k}) = E(1_A(W_{j,k})) \quad \text{a.e.} \tag{3.10}$$

(Note that the term $E(1_A(W_{j,k}))$ is independent of k.) By the Central Limit Theorem ([2])

$$\lim_{j\to\infty} E(1_A(W_{j,k})) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx. \tag{3.11}$$

It follows that

$$\lim_{j\to\infty}\lim_{N\to\infty}\frac{1}{N}\#\{W_{j,k}\in A,\ k\le N\}=\lim_{j\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}1_A(W_{j,k})=$$

$$\frac{1}{\sqrt{2\pi}}\int_A e^{-x^2/2}dx \quad \text{a.e.} \tag{3.12}$$

Now suppose $x_1, x_2, \ldots$ is a random sequence associated with the sequence $X_1, X_2, \ldots$, and for $j\ge 1,\ k\ge 1$, let

$$w_{j,k}=\frac{1}{\sqrt{j}\sigma}\sum_{i=(k-1)j+1}^{kj}(x_i-\mu).$$

Assume that the function 1_A can be computed by algorithm; (*e.g.*,this will be the case if the endpoints of A are rational numbers.) Assume also that μ and σ are computable (*i.e.*, that there are algorithms which compute arbitrarily good rational approximations of these numbers.) Then it follows easily from our definition of random sequence that

$$\lim_{j\to\infty}\lim_{N\to\infty}\frac{1}{N}\#\{w_{j,k}\in A,\ k\le N\}=\lim_{j\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}1_A(w_{j,k})=\frac{1}{\sqrt{2\pi}}\int_A e^{-x^2/2}dx \tag{3.13}$$

(*Remark*: If A has rational endpoints, then the values $1_A(w_{j,k})$ can be computed by algorithm exactly once one knows a sufficiently good rational approximation of the $w_{j,k}$, provided that $w_{j,k}$ is not an endpoint of A. This is important because **Definition 3.2** involves functions f_j which can be computed by algorithm, and we want to apply this to a case in which the f_j involve 1_A in the present example. The fact that $1_A(x)$ is not determined by knowing approximate values of x if x is an endpoint of A does not affect the truth of (3.13) due to the fact that points have probability 0 with respect to the normal distribution.)

Clearly, multidimensional (vector) analogs of (3.13) also hold if we replace A by a rectangular region with rational vertices. Again, more general regions could be utilized; we discuss this situation later.

In particular (3.13) holds in the case we have been discussing of random sequences associated with the uniform distribution on the two point set $\{0,1\}$.

The "Law of the Iterated Logarithm"

$$\limsup_{n\to\infty}\frac{\sum_{i=1}^{n}(x_i-1/2)}{\sqrt{n\log(\log n)/2}}=1 \tag{3.14}$$

also holds for random sequences associated with the uniform distribution on $\{0,1\}$ because the corresponding theorem for the associated random variables holds a.e. ([4]) and one could take

$$f_n(x_1, \ldots, x_n) = \frac{\sum_{i=1}^n (x_i - 1/2)}{\sqrt{n \log(\log n)/2}}$$

in **Definition 3.2**. Recall that (3.14) does not generally hold for "collectives" in the sense of von Mises and Church ([4]).

Recall also that there are "collectives" for which the frequency of 1's in the first n terms is always $\geq \frac{n}{2}$ ([4]). However, this clearly cannot occur for our random sequences since for the corresponding random variables X_i, if we put

$$f_n(X_1, \ldots, X_n) = \left\{ \begin{matrix} 1 & \text{if} \quad X_1 + \ldots + X_k \geq k/2, \quad \text{all } k = 1, \ldots, n \\ 0 & \text{otherwise} \end{matrix} \right\}$$

then

$$\lim_{n \to \infty} f_n(X_1, \ldots, X_n) = 0 \quad \text{a.e.}$$

(This follows from the Law of the Iterated Logarithm ([4]), for example), and then **Definition 3.2** can be applied. Notice that the condition that "the frequency of 1's in the first n terms is always $\geq \frac{n}{2}$" is a more delicate condition than the conditions on relative asymptotic frequencies that collectives must satisfy (since the former condition is one involving frequencies rather than *relative* frequencies.)

We now consider in what sense the general frequency interpretation of probabilities is a consequence of the properties of random sequences. Suppose as above that $X_1, X_2, \ldots$ are independent and identically distributed with expectation μ, variance σ^2 and distribution $dF = P_X$, and let $x_1, x_2, \ldots$ be a random sequence associated with dF. For fixed j and $k \geq 1$, let

$$\mathbf{W}_k = (X_{(k-1)j+1}, \ldots, X_{kj}) \tag{3.15}$$

and

$$\mathbf{w}_k = (x_{(k-1)j+1}, \ldots, x_{kj}). \tag{3.16}$$

Let A be a measurable subset of $\mathbb{R}^j$ and 1_A its characteristic function. Then the $1_A(\mathbf{W}_k)$ are independent and identically distributed, and by the Strong Law of Large Numbers

$$\lim_{N \to \infty} \frac{1}{N} \#\{\mathbf{W}_k \in A,\ k \leq N\} = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^N 1_A(\mathbf{W}_k) = E(1_A(\mathbf{W}_k)) = P(\mathbf{W}_k \in A) \quad \text{a.e.} \tag{3.17}$$

(The $\mathbf{W}_k$ have distribution $P_{\mathbf{W}} = (dF)^j =$ the j-fold product $dF \times \ldots \times dF$.)
We would like to conclude that the corresponding equation

$$\lim_{N \to \infty} \frac{1}{N} \#\{\mathbf{w}_k \in A,\ k \leq N\} = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^N 1_A(\mathbf{w}_k) = P(\mathbf{W}_k \in A) \tag{3.18}$$

holds for the sequence $\mathbf{w}_k$. However, in order to apply our **Definition 3.2** with f_j appropriate to the present situation, it seems necessary to make some restriction on A to ensure that the function 1_A is computable by algorithm (as was done in the discussion involving the Central Limit Theorem above.) Once we know that (3.18) holds for a reasonable collection of sets A, we will deduce that it holds for a much larger class of sets by a separate argument. We introduce some terminology. Denote the sequence $\mathbf{w}_k$ of (3.16) by $\mathbf{w}$. For a set A in $\mathbb{R}^j$, let

$$\begin{aligned} fr_N(\mathbf{w}, A) &= \frac{1}{N}\#\{\mathbf{w}_k \in A, k \leq N\} \\ \underline{fr}(\mathbf{w}, A) &= \lim \inf_{N\to\infty} fr_N(\mathbf{w}, A) \\ \overline{fr}(\mathbf{w}, A) &= \lim \sup_{N\to\infty} fr_N(\mathbf{w}, A). \end{aligned} \tag{3.19}$$

If $\underline{fr}(\mathbf{w}, A) = \overline{fr}(\mathbf{w}, A)$, then we denote the common value by $fr(\mathbf{w}, A)$. Finally, if $fr(\mathbf{w}, A) = P_{\mathbf{W}}(A)$, we call A a $\mathbf{w}$-set. It follows easily from these definitions that $\underline{fr}$, $\overline{fr}$, and fr are order preserving and finitely additive (*i.e.*, if A and B are disjoint, then $\underline{fr}(A) \leq \underline{fr}(A \cup B) = \underline{fr}(A) + \underline{fr}(B)$, and similarly for $\overline{fr}$ and fr.) Furthermore, the $\mathbf{w}$-sets are closed under taking finite disjoint unions and complements (but not countable disjoint unions or arbitrary finite unions). Now j-dimensional rectangles A with rational vertices and for which the boundary ∂A has $P_{\mathbf{W}}$ measure 0 are evidently $\mathbf{w}$-sets, since the value $1_A(\mathbf{w}_k)$ can be computed exactly given a sufficiently good rational approximation to $\mathbf{w}_k$ if $\mathbf{w}_k \notin \partial A$. Hence, finite unions of such rectangles are also $\mathbf{w}$-sets. (In this case finite unions are the same as finite disjoint unions.) The following result exhibits a much larger collection of $\mathbf{w}$-sets.

Theorem 3.20

Suppose A is a $P_{\mathbf{W}}$ measurable set and for every $\epsilon > 0$ there exist $\mathbf{w}$-sets A_1, A_2 with $A_1 \subseteq A \subseteq A_2$ and $fr(\mathbf{w}, A_2) - fr(\mathbf{w}, A_1) \leq \epsilon$. Then A is a $\mathbf{w}$-set.

Proof:

We have, for every $\epsilon > 0$,

$$P_{\mathbf{W}}(A_1) = fr(\mathbf{w}, A_1) \leq \underline{fr}(\mathbf{w}, A) \leq \overline{fr}(\mathbf{w}, A) \leq fr(\mathbf{w}, A_2) = P_{\mathbf{W}}(A_2) \leq P_{\mathbf{W}}(A_1) + \epsilon \tag{3.21}$$

Hence $fr(\mathbf{w}, A)$ exists and $= P_{\mathbf{W}}(A)$.

For reasonable discrete or continuous $P_{\mathbf{W}}$, the $\mathbf{w}$-sets comprise a large class. However, it is clear that in general not all (measurable) sets A will be $\mathbf{w}$-sets. For example, one could take A to be the set equal to the union of all the points $\mathbf{w}_k$; in this case, evidently $fr(\mathbf{w}, A) = 1$, although it might be that $P_{\mathbf{w}}(A) \neq 1$. Of course, in such a situation the function 1_A is not computable by algorithm. The frequency interpretation of $P_{\mathbf{W}}(A)$

using the sequence $\mathbf{w}_k$ is only valid for those sets A for which an "algorithmic test" of the frequency interpretation is possible (in the sense of our preceeding discussion.)

The fact that the $\mathbf{w}$-sets are plentiful provides a strong justification of the frequency interpretation of probabilities in the sense intuitively discussed by H. Cramér and mentioned earlier. Even more, as we have seen, the notion of generic and random sequences provides a model in which "all" useful probabilistic notions have an unambiguous meaning. Consider the situation of a gambler betting on the results of a sequence of coin tosses. If it is assumed that the sequence of results is a random sequence (associated with the uniform distribution on a two point set) then notions such as "the probability of a H is 1/2", "the distribution of H's is asymptotically normal", "the law of the iterated logarithm is satisfied", *etc*, have a definite meaning within the model provided by the assumption. Even the notion that if our gambler is presently losing, *his luck must eventually change* becomes a meaningful (and true) statement (this involves a straightforward generalization of the discussion above of binary sequences for which the frequency of 1's in the first n terms is always $\geq \frac{n}{2}$.) One could ask whether the sequence of outcomes is "really" a random sequence, but the question certainly has no real meaning in the context of Mathematics (or even science in general), but belongs more properly to Philosophy. (We do make some remarks about this question later.)

Generic sequences have another not unexpected property: "reasonable" functions of generic sequences are again generic. Suppose that the sequence $x_1, x_2, x_3, \ldots$ is *generic* for the sequence of random variables $X_1, X_2, X_3, \ldots$, and suppose $h(x)$ is a real function of one variable. We investigate when the sequence of $h(x_i)$ is generic for the sequence $h(X_i)$. For this to be the case, it must be that for algorithms of the type described preceeding (3.1), with

$$P(\limsup_{j\to\infty} f_j(h(X_{n_1}), \ldots, h(X_{n_j})) = \alpha) = 1 \tag{3.22}$$

we have

$$\limsup_{j\to\infty} f_j(h(x_{n_1}), \ldots, h(x_{n_j})) = \alpha \qquad (3.23).$$

This would be the case if the functions $f_j(h(x_{n_1}), \ldots, h(x_{n_j}))$ are functions $\tilde{f}_j(x_{n_1}, \ldots, x_{n_j})$ computed by an algorithm of the required type. This will clearly be the case provided $h(x)$ has the property that its value can be computed by algorithm to any degree of accuracy given a sufficiently accurate rational approximation of its argument. We might call such a real function "computable"; evidently, many familiar functions (such as $\sqrt{x}$, $\sin(x)$, e^x, $\log(x)$, *etc*) are computable in this sense. If h is computable, then one can combine an algorithm computing the $f_j(h(x_{n_1}), \ldots, h(x_{n_j}))$ in response to input of (approximations of) the $h(x_i)$ with an algorithm computing the $h(x_i)$ (approximately) to get an algorithm computing the $\tilde{f}_j(x_{n_1}, \dots, x_{n_j}) = f_j(h(x_{n_1}), \dots, h(x_{n_j}))$ in response to input of (approximations of) the x_i. It is easy to see that this line of reasoning can be extended to functions $h(x, y, ...)$ of several variables. For example, if $h(x, y)$ is a "computable" function of 2 variables, then the sequence $h(x_1, x_2), h(x_3, x_4), \dots$ is generic for $h(X_1, X_2), h(X_3, X_4), \dots$.

Many other properties of generic sequences could be discussed. However, our primary aim in the present article has been to indicate the relevance of this notion to the foundations of probability and statistics, and we discuss this further in the following section.

Section 4: Complements

As we have discussed, it is difficult to assign an operational meaning to most probabilistic notions. This is due to the fact that real processes are usually thought of as being intrinsically deterministic. For events that only occur once and for which the idea of repetition is not really meaningful (*e.g.* the event that a particular inmate is released from jail, or the event that a cure for cancer is found, *etc*), a probabilistic interpretation seems necessarily subjective. However, for other processes, such as tossing a coin, which are repeated, or *whose repetition can be envisaged*, some probabilities (*e.g.* the probability of obtaining Head when tossing a coin) can be given a frequency interpretation. In order for such an interpretation to be directly connected to what is actually occurring, the frequencies involved should refer to an actual sample sequence. (Some of these involve relative frequencies, but as we have seen, others are more delicate and involve absolute frequencies.)

However, there are many (often uncountably many) events whose probabilities should be interpretable as frequencies, and it often happens that the sample sequences satisfying all these conditions have probability zero. This was a difficulty in von Mises's approach; in addition, his principle of the impossibility of advantageous gambling systems expressed in terms of "place selections" (selecting subsequences depending only on what has occurred in the past) really did not make good sense mathematically. Church's use of computable functions to define the place selections was just what was needed, but did not go far enough (recall Ville's objections.) Martin-Löf's idea of *effective statistical tests of randonmess* overcomes such objections but, as formulated, only in the case of binary sequences. The notion of *generic sequence* of the present article resolves all these problems and in a quite general setting. When we are concerned with some process for which repetition is sensible, then our model for for the outcomes is a generic sequence associated with some abstract sequence of random variables. Then, as has been indicated previously, all algorithmically described probabilistic properties which are "generic" (*i.e.* hold with probability 1 for the related abstract process) hold for our model in some concrete sense (which might involve "frequencies", but for which an abstract notion of "probability" is no longer involved.)

E.g., for fair coin tossing, our model is a binary sequence for which the relative frequency of Heads is really $\frac{1}{2}$, the frequency of Heads in the first n tosses will sometimes be significantly greater and sometimes significantly less that $\frac{n}{2}$, *etc*, *etc.* As remarked earlier, this implies that a gambler making even money bets on the individual outcomes might be losing significantly for a while, but (assuming the coin is "fair"), he can be sure that eventually he will be significantly ahead (assuming that he lives long enough and his bankroll isn't depleted forcing a halt.) Of course, our supposition that the sequence of outcomes is, in fact, a generic sequence is a supposition which may or may not be true, but this is the case with models generally (including all the "laws" of physics.) A difference from the usual (axiomatic) probabilistic interpretation is that what is "true" for our model would be said to hold with probability 1 in the axiomatic interpretation, but this distinction has no real significance in practice (since presumably we will not observe the future occurrence of events whose probability we now believe is 0.) What is significant here is the fact that there is no need to resolve the dichotomy of the deterministic *vs* probabilistic; the process giving rise to a supposed generic sequence might well be deterministic (although not "algorithmic".) For example, in the case of a person tossing a coin, one might envision the results as determined by some straightforward calculation involving the forces acting on the coin, but, of course, this is not at all the case. The results are determined by the totality of effects (initial conditions and forces) in the vicinity (which may be quite extensive) of the experiment, including gravity, atmospheric effects, the emotions of the experimenter, relativistic and quantum mechanical effects, *etc.* It is, therefore, not at all clear that the sequence of results obtained by repeated tossing can be produced by some algorithm; indeed, this seems rather unlikely. It is also the case that there is no practical way to test whether a sequence occurring in this way is actually generic (random), since any initial segment of a generic sequence can be completely arbitrary. For this reason, a hypothesis that some sequence occurring naturally is generic usually can't be "disproved", although we may have some suspicion one way or the other. Our model for the outcomes of successive spins of a roulette wheel may involve the assumption that the sequence obtained is random (associated with the obvious sequence of independent random variables.) If the sequence appears predictable in some aspect (*e.g.* some number occurs more than expected), this may be cause to reject the model, but a definite conclusion probably can't be reached. As we know, the initial part of a random sequence may seem predictable, even though this predictability disappears after a sufficiently long time. This is exactly what makes the play so seductive to some; even with an "unbiased" wheel, it is possible to be ahead for quite a long time (although not indefinitely if the model is correct.) In any case, the point is that the sequence of outcomes is deterministic, but this seems compatible with the assumption of randomness (which now has a definite meaning.)

It is natural to wonder whether the result of some physical process, *e.g.* tossing a coin repeatedly, might actually produce a random sequence of the type defined above. Undoubtedly, this is not really meaningful in a precise way; even the notion of an *infinite* repetition may have no real meaning. However, we might understand the question as an inquiry concerning the usefulness of the model and assign meaning in this way. In the case of coin tossing, it seems reasonable that, even if the sequence of outcomes is considered a random sequence, the associated probability of obtaining a head is not .5, but rather some other value possibly quite close to .5. Of course, if the difference is sufficiently small, it might be unnoticeable and hence irrelevant. It seems just as likely, however, that the sequence obtained is not random in the sense we have considered; perhaps the relative frequency of heads does not have any limiting value, but rather fluctuates indefinitely. Nevertheless, it might be that the range of this fluctuation is eventually so small that the notion of a limiting value is appropriate in the model considered. As remarked above, the connection between models and reality is generally problematical in this way.

We note, finally, that in the framework of Statistics it is the notion of random (rather than generic) sequence that is most relevant, since the mathematical techniques employed generally apply to collections of independent, identically distributed, random variables. In this regard, it is should be realized that in many cases, **all** of the mathematical implications of the usual axiomatic framework follow from the properties of a single random sequence, since, as discussed above, the joint finite distributions of the associated random variables are determined by the properties of a random sequence. (See (3.17), (3.18) and the following discussion. Recall that (3.18) only was proved for $\mathbf{w}$-sets A, but this is sufficient to determine $P(\mathbf{W}_k \in A)$ for all measurable A for many distributions.) In these cases, the model provided by the sequence of randon variables is equivalent to that provided by a random sequence.

References

[1] A. Church, *The Concept of a Random Sequence*, Bull. Amer. Math. Soc., 46, 130-135, 1940.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 18th printing, 1991.

[3] J.L. Doob, *Note on Probability*, Annals of Mathematics, Vol. 37, No. 2, 363-367, April, 1936.

[4] Ming Li, Paul Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York, 1993.

[5] P. Martin-Löf, *On the Concept of a Random Sequence*, Theory Probability Appl., 11, 177-179, 1966.

[6] L.J. Savage, *The Foundations of Statistics*, Dover, New York, 1972.

[7] Richard von Mises, *Probability, Statistics, and Truth*, Dover, New York, 1957.