THE FOUNDATIONS OF PROBABILITY

P. R. HALMOS, Syracuse University

1. Introduction. Probability is a branch of mathematics. It is not a branch of experimental science nor of armchair philosophy, it is neither physics nor logic. This is not to say that the experimenter and the philosopher should not discuss probability from their points of view. They should, and they do. The situation is analogous to that in geometry. No one denies that the physicist and the philosopher have made valuable contributions to our understanding of the space concept, nor, in spite of this, that geometry is a rigorous part of modern mathematics.

Like Euclidean geometry, and for that matter like most mathematical theories, probability has four aspects: axiomatization, development, coordinatization, and application. We proceed to explain our use of these words.

"Axiomatization" is clear. We all know that the study of geometry begins with a list of undefined terms and a list of postulates. It is important in this connection to remember two facts. First: the selection of the list of terms and postulates is not entirely arbitrary, but is derived only after a thorough examination of our intuitive notions of the subject. Second: the selection of terms and postulates is not uniquely determined. When several different axiomatizations of the same subject exist then only extra mathematical considerations, such as practical convenience or personal prejudice, can lead us to prefer one among the many. The greater part of this paper is devoted to a prepostulational examination of probability. The axiomatic system to which this examination leads is not the only possible approach to probability, but it is the approach which has been adopted by the majority of workers in this field.

By "development" we mean simply the main part of the theory, the definitions and theorems which chiefly occupy the professional mathematician. "Coordinatization" is a general process the most familiar instance of which is the proof of the equivalence of the synthetic and analytic aspects of Euclidean geometry. The isomorphism of a finite group to a group of permutations and the representation of an algebra by matrices are further examples of this process. Properly speaking coordinatization is just one of the theorems belonging to development, but a theorem of such fundamental implications that it effects basic changes in the appearance, methods, and results of the entire theory.

The hardest philosophical problem in geometry as well as in probability is the problem of "application." Do the theorems derived from the postulates reflect any light on the physical world which suggested them, and if so, how and why?

The purpose of this paper is exposition, exposition intended to convince the professional mathematician that probability is mathematics. To this end we shall discuss the four features just enumerated. The paper contains almost no proofs, very few precise definitions and theorems, and many heuristic derivations. Despite however the small number of rigorous statements, they form the

foundation on which the remainder is built. For the convenience of the reader they are italicized. If these italicized statements are lifted from their context and read consecutively, they will furnish at least a partial answer to the question "what is probability?"

2. Boolean algebra. The principal undefined term in probability theory is "event." Intuitively speaking an event is one of the possible outcomes of some physical experiment.

To take a rather popular example consider the experiment of rolling an ordinary six-sided die and observing the number $v(=1, 2, 3, 4, 5, \text{or } 6)$ showing on the top face of the die. "The number v is even"—"it is less than 4"—"it is equal to 6"—each such statement corresponds to a possible outcome of the experiment. From this point of view there are as many events associated with the experiment as there are combinations of the first six positive integers taken any number at a time. If for the sake of aesthetic completeness and later convenience we consider also the impossible event, "the number v is not equal to any of the first six positive integers," then there are altogether 2^6 admissible events associated with the experiment of the rolling die. For the purpose of studying this example in more detail let us introduce some notation. We write $\{246\}$ for the event "v is even," $\{123\}$ for "v is less than 4," and so on. The impossible event and the certain event $(=\{123456\})$ deserve special names: we reserve for them the symbols o and e respectively.

Everyday language concerning events uses such phrases as these: "two events a and b are incompatible or mutually exclusive," "the event a is the opposite of the event b or complementary to b," "the event a consists of the simultaneous occurrence of b and c," "the event a consists of the occurrence of at least one of the two events b and c." Such phrases suggest that there are relations between events and ways of making new events out of old that should certainly be a part of their mathematical theory.

The notion of complementary event is probably closest to the surface. If a is an event we denote the complementary event by a': an experiment one of whose outcomes is a will be said to result in a' if and only if it does not result in a. Thus if $a=\{246\}$ then $a'=\{135\}$. We may also introduce combinations of events suggested by the logical concepts of "and" and "or." With any two events a and b we associate their "join" $a\cup b$ (also called union or sum and often denoted by $a+b$), and their "meet" $a\cap b$ (or intersection or product, often denoted by ab). Here $a\cup b$ occurs if and only if at least one of the two events a or b occurs, while $a\cap b$ occurs if and only if both a and b occur. Thus if $a=\{246\}$ and $b=\{123\}$ then $a\cup b=\{12346\}$ and $a\cap b=\{2\}$.

The operations a', $a\cup b$, and $a\cap b$ satisfy some simple algebraic laws. It is clear for example that both the expressions $a\cup b$ and $a\cap b$ are independent of the order of the terms (commutative law), and that neither of the expressions $a\cup b\cup c$ and $a\cap b\cap c$ depends on the order in which the two indicated operations are performed (associative law). These facts are intuitively obvious from the verbal definition of the operations and are easily verified in any finite case such

as the rolling die. There are many other similar identities satisfied by these methods of combining events: the following is a list of the most important ones.

$$o' = e \qquad (a')' = a \qquad e' = o$$

$$(a \cap b)' = a' \cup b' \qquad (a \cup b)' = a' \cap b'$$

$$a \cap a' = o \qquad a \cup a' = e$$

$$o \cap a = o \qquad o \cup a = a$$

$$e \cap a = a \qquad e \cup a = e$$

$$a \cap b = b \cap a \qquad a \cup b = b \cup a$$

$$(a \cap b) \cap c = a \cap (b \cap c) \qquad (a \cup b) \cup c = a \cup (b \cup c)$$

$$a \cap (b \cup c) = (a \cap b) \cup (a \cap c) \qquad a \cup (b \cap c) = (a \cup b) \cap (a \cup c)$$

A system B of elements o, a, b, $\cdots$, e in which operations a', $a \cup b$, and $a \cap b$ are defined in such a way that each of the above list of identities is satisfied is called a "Boolean algebra." For the traditional theory of probability, concerned with simple gambling games such as the rolling die, in which the total number of possible events is finite, the above heuristic reduction of events to elements of a Boolean algebra is adequate. For situations arising in modern theory and practice, and even for the more complicated gambling games, it is necessary to make an additional assumption. This assumption, in descriptive terms, is that the operations $\cup$ and $\cap$, assumed defined for two elements and immediately extended by mathematical induction to any finite number, should make sense also for an infinite sequence. In other words it is desirable to have an interpretation for symbols such as $a_1 \cup a_2 \cup \cdots$ and $a_1 \cap a_2 \cap \cdots$. In order to phrase precisely this assumption of infinite operations it is necessary to use a few simple facts from the theory of Boolean algebras.

If a and b are any two elements of the Boolean algebra B which satisfy the relation $a \cup b = b$ (or the equivalent relation $a \cap b = a$) we shall write $a \subset b$ and say that "a is smaller than b" or "a is contained in b" or "a implies b." The intuitive interpretation of this relation is as follows: the event a implies the event b, or is contained in the event b, if the occurrence of a is a sub-case of the occurrence of b. Thus in the example of the die $\{123\} \subset \{1234\}$ and "$v = 2$" $\subset$ "v is even." The technical significance of the relation $\subset$ is that the operations $\cup$ and $\cap$ may be defined in terms of it. For example $a \cup b$ is the smallest of all elements which contain both a and b. In more detail: given a and b, consider all c's for which both $a \subset c$ and $b \subset c$. The assertion concerning $a \cup b$ is two fold: first, $a \cup b$ is an admissible c, and second, for any admissible c we have $a \cup b \subset c$. As an example consider $a = \{12\}$ and $b = \{24\}$. The elements $\{1234\}$, $\{1246\}$, $\{12456\}$, $\cdots$ all have the property of containing both a and b. However the element $\{124\}$, which also has that property, is smaller than any other such element, and it is in fact true that $\{12\} \cup \{24\} = \{124\}$.

Motivated by the relation between $\cup$ and $\subset$ we now proceed as follows. Let B be a Boolean algebra. If for every infinite sequence $a_1, a_2, \cdots$ of elements of

B there exists among the elements containing all the a_n a smallest one, say a, we say that B is a σ-algebra and we write $a = a_1 \cup a_2 \cup a_3 \cup \cdots$. Not every Boolean algebra is a σ-algebra; the assumption that B is one (the hypothesis of countable additivity) is an essential restriction.

Perhaps an example, though a somewhat artificial one, might illustrate the need for the added assumption. Suppose that a player determines to roll a die repeatedly until the first time that the number showing on top is 6. Let a_n be the event that the first 6 appears only on the nth roll. The event $a = a_1 \cup a_2 \cup a_3 \cup \cdots$ occurs if and only if the game ends in a finite number of rolls. The occurrence of the opposite event a' is at least logically (even if not practically) conceivable and it seems reasonable to want to include a discussion of it in a general theory of probability. Numerous examples of this kind together with some rather deep lying technical reasons justify therefore the following statement.

The mathematical theory of probability consists of the study of Boolean σ-algebras.

This is not to say that all Boolean σ-algebras are within the domain of probability theory. In general statements concerning such algebras and the relations between their elements are merely qualitative: probability theory differs from the general theory in that it studies also the quantitative aspects of Boolean algebras. In the next section we shall describe and motivate the introduction of numerical probabilities.

3. Measure algebra. When we ask "what is the probability of a certain event?" we expect the answer to be a number, a number associated with the event. In other words probability is a numerically valued function P of events a, that is of elements of a Boolean σ-algebra B, $P = P(a)$. On intuitive and practical grounds we demand that the number $P(a)$ should give information about the occurrence habits of the event a. If in a large number of repetitions of the experiment which may result in the event a we observe that a actually occurs only a quarter of the time (the remaining three quarters of the experiments resulting therefore in a') we may attempt to summarize this fact by saying that $P(a) = 1/4$. Even this very rough first approximation to what is desired yields some suggestive clues concerning the nature of the function P.

If, to begin with, $P(a)$ is to represent the proportion of times that a is expected to occur, then $P(a)$ must be a positive real number, in fact a number in the unit interval $0 \leqq P(a) \leqq 1$. The extreme value 0 has a special significance. Since the impossible event o will never occur, it is clear that we must write $P(o) = 0$. Conversely however if an event a refuses ever to occur, we are tempted to declare its occurrence impossible and thus from the relation $P(a) = 0$ to deduce $a = o$. The other extreme value of $P(a)$ has of course a similar interpretation: $P(a) = 1$ if and only if $a = e$.

The relation between proportion and probability has further consequences. Suppose that a and b are mutually exclusive events—say $a = \{1\}$ and $b = \{246\}$ in the example of the die. (In the algebraic theory mutually exclusive events correspond to "disjoint" elements of the Boolean algebra B, that is to elements

a and b for which $a \cap b = o$.) In this case the proportion of times that the join $a \cup b (= \{1246\}$ for the example) occurs is clearly the sum of the proportions associated with a and b separately. If an ace shows up one-sixth of the time and an even number half the time, then the proportion of times in which the top face is either an ace or an even number is $\frac{1}{6}+\frac{1}{2}$. It follows therefore that the function P cannot be completely arbitrary—it is necessary to subject it to the condition of additivity, that is to require that if $a \cap b = o$ then $P(a \cup b)$ should be equal to $P(a)+P(b)$.

We are now separated from the final definition of probability theory only by a seemingly petty (but in fact very important) technicality. If $P(a)$ is an additive function of the sort just described on a Boolean σ-algebra B, and if $a_1, a_2, \cdots, a_n$ is any finite set of pairwise disjoint elements of B (this means that for $i \neq j$, $a_i \cap a_j = o$) then it's easy to prove by mathematical induction that $P(a_1 \cup a_2 \cup \cdots \cup a_n) = P(a_1)+P(a_2)+\cdots+P(a_n)$. If however $a_1, a_2, a_3, \cdots$ is an infinite sequence of pairwise disjoint elements then it may or may not be true that $P(a_1 \cup a_2 \cup a_3 \cup \cdots) = P(a_1)+P(a_2)+P(a_3)+\cdots$. The general condition of countable (that is, finite or enumerably infinite) additivity is a further restriction on the probability measure P—a restriction without which modern probability theory could not function. It is a tenable point of view that our intuition demands infinite additivity just as much as finite additivity. At least however infinite additivity does not contradict any of our intuitive ideas and the theory built on it is sufficiently far developed to assert that the assumption is justified by its success. We shall therefore adopt this assumption as our final postulate.

Numerical probability is a measure function, that is a finite, nonnegative, and countably additive function P of elements in a Boolean σ-algebra B, such that if the null and unit elements of B are o and e respectively then $P(a) = 0$ is equivalent to $a = o$ and $P(a) = 1$ is equivalent to $a = e$.

In the next section we shall discuss a general method of constructing examples of probability measures.

4. Measure space. Let $\omega_j (j = 1, \cdots, 6)$ be the point on the real axis whose directed distance from the origin is j, and let Ω be the set whose elements are these six points. Consider the system B^* of all subsets of Ω. (The empty set o and the full set $e = \Omega$ are counted as belonging to B^*.) With any element a of B^* (that is, with any subset of Ω) we may associate the complementary element (set) consisting of exactly those points ω_j which do not belong to a. Similarly with any two subsets a and b of Ω we may associate their union (the set of points belonging to either a or b or both), and their intersection (the set of points belonging simultaneously to a and b). It is easy to verify that under the operations of complementation (a'), formation of unions ($a \cup b$), and formation of intersections ($a \cap b$), the system B^* forms a Boolean algebra, in fact, though somewhat vacuously, a σ-algebra. Suppose moreover that for each $j = 1, \cdots, 6$, p_j is a positive number such that $p_1 + \cdots + p_6 = 1$. Then we may define $P(a)$ for any subset a of Ω, to be the sum of those p_j whose ω_j belongs to a. Thus if

$a=\{135\}$ then $P(a)=p_1+p_3+p_5$; if $a=o$ then $P(a)=0$. The function P and the algebra B^* satisfy all the assumptions of probability theory and the reader has doubtless recognized that this B^* and P were implicit in our earlier discussion of the rolling die. It is often customary on philosophical and practical grounds to discuss only the case $p_1=\cdots=p_6=\frac{1}{6}$. We shall say a word about this special case later; for the moment it is sufficient to point out that any other choice of the p_j furnishes an equally acceptable probability structure and does in fact constitute the mathematical theory of some carefully loaded die.

The above example of a Boolean algebra can be generalized: we attempt next to obtain a similar but more geometrical example. For this purpose we again choose a set Ω, but, instead of a finite set, we choose a set with infinitely many points, in fact all the points of a continuum. To be specific let us choose for Ω the points ω of a square of unit area in the Cartesian plane. In analogy with the preceding example we consider the system B^* of all subsets of Ω and define complement, union, and intersection as before. Once more B^* is a Boolean σ-algebra; it is not however the one on which we shall base our probability theory. (It can be shown that it is not possible to define a probability measure P with the desired properties on B^*.) We shall instead consider a certain subsystem (sub-algebra) of B^*, constructed as follows:

We begin with the system R of all rectangles contained in Ω (where for the sake of definiteness we consider closed rectangles, that is sets consisting of the interior plus the perimeter of a rectangle). The system R is not closed under the Boolean operations: in general not even a finite (let alone a countably infinite) union or intersection of rectangles is itself a rectangle, and similarly the complement of a rectangle isn't one. We have therefore to enlarge the system R to a system R' including all complements and countable unions and intersections of elements of R. It turns out that even this is not enough: R' is still not a Boolean algebra, and the extension process has to be continued. If however the extension process is continued sufficiently (and this happens to mean transfinitely) often, we reach eventually a Boolean σ-algebra B of subsets of Ω. (The algebra B is important in analysis: sets of B are called the Borel sets of the square.)

We face next the task of defining P. For those familiar with the theory of Lebesgue measure it will suffice to say that we define $P(a)$, for each a in B, to be the Lebesgue measure of the set a. It is not difficult to get an intuitive idea of how P is defined. If a is a rectangle (that is an element of R) we define $P(a)$ to be the area of a. If a is an element of R' we proceed to determine $P(a)$ in accordance with the requirement of countable additivity. Thus for example if b is the complement of a rectangle a, we write $P(b)=1-P(a)$, and if b is the union of a finite or infinite sequence of disjoint rectangles $a_1, a_2, \cdots$ we define $P(b)=P(a_1)+P(a_2)+\cdots$. By repeating this extension process ad transfinitum we succeed eventually in defining $P(a)$ for every a in B.

There is an objection to the construction just described. If the set a consists of a single point then it is intuitively obvious (and follows easily from the rigorous definition of P) that $P(a)$ (=the area of a) is zero. More generally if a

consists of any finite or enumerably infinite set of points we still have $P(a)=0$, and it is even possible (if for example a is a line segment) to have $P(a)=0$ for sets a containing uncountably many points. This definitely contradicts our explicitly formulated axiom that $P(a)=0$ should happen if and only if $a=o$. The customary way to get around this difficulty is by redefining the notion of equality that occurs in the equation $a=o$. It is proposed that we agree to consider as identical two subsets of Ω whose difference has probability zero. (In technical language, we consider, instead of the sets a, equivalence classes of sets modulo the class of sets of probability zero.) Through this agreement we are committed in particular to identifying any set of probability zero with the empty set o, and it follows therefore that in the reduced algebra B (that is, the algebra obtained from B by making the suggested identifications) all the axioms of probability are valid.

The long and tortuous process just described is very general. If Ω is any space (such as an interval or a cube) on a certain σ-algebra B of subsets of which a countably additive measure P is defined (such as length or volume), subject only to the restriction that the measure of all Ω is equal to 1, we obtain from B and P a system satisfying all the axioms of probability theory by the process of identification according to sets of measure zero. Thus there are as many probability systems as there are examples of "measure spaces."

The reason for the introduction of measure spaces into a discussion of probability theory is not merely to give examples. It can in fact be shown that the two theories (measure and probability) are coextensive. More precisely:

If B is any Boolean σ-algebra and P a probability measure on B, then there exists a measure space Ω such that the system B is abstractly identical with an algebra of subsets of Ω reduced by identification according to sets of measure zero, and the value of P for any event a is identical with the values of the measure for the corresponding subsets of Ω.

Hence measure is probability and probability is measure and, in virtue of the theorem just stated, the entire classical theory of measure and integration may be and has been carried over and used to give rigorous proofs of probability theorems.

5. Measure vs. probability. Having discussed the extent to which probability and measure are the same, we now dedicate a few words to describing the extent to which they are different. One feature that differentiates the two theories is that in the general theory of measure it is usual to admit the possibility that the measure of the entire space is infinite. This possibility is not admissable in probability theory. As long, however, as the measure of the whole space is finite it is always possible to introduce a scale factor which makes it equal to 1, and hence it is always possible to think of it (even if somewhat artificially) as a "probability space." Thus for example the language and notation of probability may be and have been used in such seemingly widely separated parts of mathematics as ergodic theory, topological groups, and integral geometry.

Even however if the infinite case is ruled out, it is a conspicuous fact that most theorems in which the word measure is used (rather than the word probability) have a very different appearance from the theorems of probability theory. The best way to explain the difference between measure and probability is to liken it to the difference between analytic and synthetic geometry. It isn't stretching a point too far to say that the representation of a probability algebra by a measure space is similar to the introduction of coordinates into geometry. Synthetic and analytic geometry are of course abstractly identical in the sense that any theorem in the one domain may be stated and proved in the language and machinery of the other—may be, but isn't. The theorems in the two fields differ in their intuitive content. It is natural to discuss linear transformations in analytic geometry and the nine point circle in synthetic geometry—and even though the interchange is possible, it isn't desired. The abstract identity of the two fields is however an extremely useful fact, exploited mostly by the synthetic side which often finds it convenient to lean on the analytic crutch. Similarly, probability is measure, and research in the field would be very greatly hampered if we were not permitted to use this analytic crutch—but the notions suggested by probability, the notions which are important and intuitive and natural inside the field, appear sometimes extremely special and artificial in the frame work of general measure theory.

In this section and the preceding ones we have treated axiomatization and coordinatization. We proceed now to development. In the following sections we shall define the basic concepts of probability theory, and discuss in particular those which serve in the sense described above to give to probability its distinguishing flavor.

6. Independent events. In order to motivate the definitions of the concepts to be studied in the sequel we return to the example of the die. For simplicity we make the classical assumption that any two faces are equally likely to turn up and that consequently the probability of any particular face showing is $\frac{1}{6}$. Consider the events $a = \{246\}$ and $b = \{12\}$. The first notion we want to introduce, the notion of conditional probability, can be used to answer such questions as these: "what is the probability of a when b is known to have occurred?" In the case of the example: if we know that v is less than 3, what can we say about the probability that v is even? The adjective "conditional" is clearly called for in the answer to a question of this type: we are evaluating probabilities subject to certain preassigned conditions.

To get a clue to the answer consider first the event $c = \{2\}$ and ask for the conditional probability of a, given that c has already occurred. The intuitive answer is perfectly clear here, and is independent as it happens of any such numerical assumptions as the equal likelihood of the faces. If v is known to be 2 then v is certainly even, and the probability must be 1. What made the answer easy was the fact that c implied a. The general question of conditional probability asks us to evaluate the extent (measured by a numerical probability or propor-

tion) to which the given event b implies the unknown event a. Phrased in this way the question almost suggests its own answer: the extent to which b is contained in a can be measured by the extent to which a and b are likely to occur simultaneously, that is by $P(a \cap b)$. Almost—not quite. The trouble is that $P(a \cap b)$ may be very small for two reasons: one is that not much of b is contained in a, and the other is that there isn't very much of b altogether. In other words it isn't merely the absolute size of $a \cap b$ that matters: it's the relation or proportion of this size to the size of b that's relevant.

We are led therefore to define the conditional probability of a, given that b has occurred, in symbols $P_b(a)$, as the ratio $P(a \cap b)/P(b)$. For $a = \{246\}$ and $c = \{2\}$ this gives the answer we derived earlier, $P_c(a) = 1$; for $a = \{246\}$ and $b = \{12\}$ we get the rather reasonable figure $P_b(a) = \frac{1}{2}$. In other words if it's known that v is either 1 or 2 then v is even or odd (that is equal to 1 or equal to 2) each with probability $\frac{1}{2}$.

Consider now the following two questions: "b happened, what is the chance of a?" and simply "what is the chance of a?." The answers of course are $P_b(a)$ and $P(a)$ respectively. It might happen, and does in the example given above, that the two answers are the same, that in other words knowledge of b contributes nothing to our knowledge of the probability of a. It seems natural in this situation to use the word "independent": the probability distribution of a is independent of the knowledge of b. This motivates the precise definition: two events a and b are independent if $P_b(a) = P(a)$. The definition is transformed into its more usual form and at the same time gains in symmetry if we recall the definition of $P_b(a)$. In symmetric form: a and b are independent in the sense of probability (statistically or stochastically independent) if and only if $P(a \cap b) = P(a)P(b)$.

7. Repeated trials. Suppose next that we wish to make two independent trials of the same experiment—say, for example, to roll an honest die twice in succession. We shall presently exploit the precise definition of independence to clarify the notion of independent trials; first however it's worth while to remark on the intuitive content of the concept. Suppose that in a crude attempt to even things up we resolve on the following procedure: if the first die shows an even number we choose for the second experiment a die on which all the numbers are odd, and vice versa. The two experiments are not independent of each other in this case: whereas the a priori probability of getting an even number with the second die is $\frac{1}{2}$, the conditional probability of getting an even number with the second die, given that the first one showed an odd number, is one. We say that the two experiments are performed independently of each other only if the conditions under which the second experiment is to be performed are unaffected by the outcome of the first experiment.

If an experiment consists of two rolls of a die we don't expect the reported outcome of the experiment to be a number v, but rather a pair of numbers (v_1, v_2). The measure space Ω associated with the two-fold experiment consists

not of 6 but of 36 points. (It is convenient to imagine these points laid out along the regular pattern of a 6×6 square.) The problem is to determine how the probability is distributed among these points. For a clue to the answer consider the events $a=$"$v_1<3$" and $b=$"$v_2<4$." We have $P(a)=\frac{1}{3}$ and $P(b)=\frac{1}{2}$; hence if we interpret the independence of the trials to mean the independence of any two events such as a and b we should have $P(a\cap b)=\frac{1}{6}$. If in the suggested diagram for the measure space associated with this discussion we encircle the points belonging to $a\cap b$ we get the following figure.

v_2 \ v_1	1	2	3	4	5	6
1	⊙	⊙	·	·	·	·
2	⊙	⊙	·	·	·	·
3	⊙	⊙	·	·	·	·
4	·	·	·	·	·	·
5	·	·	·	·	·	·
6	·	·	·	·	·	·

We see therefore that the formula $P(a\cap b)=P(a)P(b)$ appears analogous to the fact that the area of a rectangle is the product of the lengths of its sides.

We say therefore, if the analytic description of an experiment is given by a measure space Ω with a Boolean σ-algebra B of subsets on which a probability measure P is defined, that the analytic description of the experiment consisting of two independent trials of the given experiment is as follows. The space of points ω is replaced by the space of pairs of points (ω_1, ω_2) (the so called product space $\Omega\times\Omega$), B is replaced by the Boolean σ-algebra generated by the "rectangular" sets of the form $\{\omega_1$ is in a_1, ω_2 is in $a_2\}$ where a_1 and a_2 belong to B, and the probability measure on this space of pairs is determined by the requirement that its value for rectangular sets of the kind described should be given by the product $P(a_1)P(a_2)$. The ideas involved in this procedure are not essentially original nor characteristic of probability theory: they are the same as the ideas involved in defining the area of plane sets in terms of the length of linear sets. There is of course a theorem hidden in this definition—a theorem which asserts that a probability measure satisfying the stated product requirement indeed exists and is in fact uniquely determined by this requirement.

What we can do once, we can do again. Just as two repetitions of an experiment gave rise to ordered pairs (ω_1, ω_2), similarly any finite number of repetitions (say n) give rise to the space of ordered n-tuples $(\omega_1, \omega_2, \cdots, \omega_n)$, with a multiplicatively determined probability measure. The procedure can be extended also to infinity: the analytic model of an infinite sequence of independent repetitions of an experiment is a measure space Ω whose points ω are infinite sequences $\{\omega_1, \omega_2, \omega_3, \cdots\}$. Even if an actually infinite sequence of repetitions of an experiment is practically unthinkable, there is a point in considering the infinite dimensional space Ω. The point is that many probability statements are asser-

tions concerning what happens in the long run—assertions which can be made precise only by carefully formulated theorems concerning limits. Hence even if practice yields only approximations to infinity, it is the infinite sequence space Ω that is the touchstone whereby the mathematical theory of probability can be tested against our intuitive ideas. The first and most important such long run statement is described in the following paragraphs.

Suppose that an experiment is capable of producing an event a with probability p, and suppose that an infinite sequence of independent trials of this experiment is performed. We consider therefore the space of all sequences $\omega = \{\omega_1, \omega_2, \omega_3, \cdots\}$ where for each n, ω_n may or may not belong to a. Once the experiments have been performed so that we are given a particular point ω we may start asking numerical questions. We may ask for example: out of the first n trials of the basic experiment how many resulted in a? This means: out of the first n coordinates $\omega_1, \omega_2, \cdots, \omega_n$ of ω how many belong to a? The answer to this question depends obviously on n and just as essentially on the particular sequence ω—let us denote it by $m_n(\omega)$.

Now what does out intuition say? The usual statement (one which we have already exploited in our heuristic derivation of the notion of probability) is that the ratio of the number of successes to the total number of trials should be approximately equal to the probability of the event being tested. In our notation this seems to mean that for large n the ratio $m_n(\omega)/n$ should be close to the constant $p = P(a)$. The question arises: for which ω's should this be true? Not surely for all of them. For the sequence space Ω contains sequences none of whose coordinates belong to a, and for such a sequence ω, $m_n(\omega)$ is zero for all n. The best that we have a right to demand is that the ω's for which our statement is not true should be equivalent to the empty set of ω's in the sense of probability—that is that their totality should have probability zero. And this is true.

To sum up: we have just derived the statement (not the proof) of the most important special case of the so called strong law of large numbers. In mathematical language the assertion of this law is that as $n \to \infty$, $\lim m_n(\omega)/n$ exists and is equal to $p (= P(a))$ except for a set of ω's of measure zero. In more classical terms: it is almost certain that the "success ratios" converge to the probability of the event being tested.

8. Random variables. In order to gain a more thorough understanding of the law of large numbers and at the same time to introduce the language in which most of the theorems of probability theory are stated, we proceed to discuss the notion of a random variable.

"A random variable is a quantity whose values are determined by chance." What does that mean? The word "quantity" is meant to suggest magnitude—numerical magnitude. Ever since rigor has come to be demanded in mathematical definitions it has been recognized that the word "variable," particularly a variable whose values are "determined" somehow or other, means in precise language a function. Accordingly a random variable is a function: a function

whose numerical values are determined by chance. This means in other words that a random variable is a function attached to an experiment—once the experiment has been performed the value of the function is known. The spatial model of probability is extremely well adapted to making this notion still more precise. If the analytic correspondent of an experiment is a measure space Ω then any possible outcome of the experiment is by definition represented by a point ω in this space. Hence a function of outcomes is a function of ω's: a random variable is a real valued function defined on a probability space Ω.

The preceding sentence does not yet constitute our final definition of a random variable. For suppose that $x=x(\omega)$ is a function on the space Ω. We shall call x a random variable only if probability questions concerning the values of x can be answered. An example of such a question is: what is the probability that x is between α and β? In measure theoretic language: what is the measure of the set of those ω's for which the inequality $\alpha \leqq x(\omega) \leqq \beta$ is satisfied? In order for such questions to be answerable it is necessary and sufficient that the sets that occur in them belong to the basic σ-algebra B of Ω. A function $x(\omega)$ for which this is true for every interval (α, β) is called "measurable." Accordingly we make the following definition:

A random variable is a measurable function defined on a measure space with total measure 1.

Instances of random variables can be found even in that part of our discussion which preceded their definition. The quantity v associated with the rolling die is an example, as are also the quantities v_1 and v_2 associated with the two fold repetition of this experiment. To obtain some further examples, consider any fixed event a which may result from an experiment and let the random variable x be the number of times that a actually occurs. If the experiment is performed only once then x has only two possible values: 1 if a occurs and 0 otherwise. More generally if the experiment is repeated n times the random variable x becomes the function $m_n(\omega)$ introduced in the discussion of the law of large numbers.

9. Expectation, variance, and distribution. Let us consider in detail the random variable v associated with an honest die. The possible values of v are the first six positive integers. The arithmetic mean of these values, that is the number $(1+\cdots+6)/6$, is of considerable interest in probability theory. It is called the average, or mean value, or expectation of the random variable v and it is denoted by $E(v)$. If the die is loaded so that the probability p_j associated with j is not necessarily $\frac{1}{6}$ then the arithmetic mean is replaced by a weighted average: in this case $E(v)=1\cdot p_1+\cdots+6\cdot p_6$. It is well known that the analogs of such weighted sums in cases where the number of values of the function (random variable) need not be finite are given by integrals. The kind of integral that enters into probability theory is similar in every detail to the Lebesgue integral and we shall not reproduce its definition here.

If the measurable function $x(\omega)$ is integrable then its expectation $E(x)$ is by definition the value of its integral extended over the entire domain Ω.

As a useful though extremely special case we mention that if x is a counting variable of the sort mentioned in the preceding paragraph ($x=1$ if a certain even a occurs and $x=0$ otherwise) then $E(x)=P(a)$.

It is obviously of interest to ask not only what is the expected value of a random variable x but also how closely the values of x are clustered about its expected value. The customary measure of clustering of a random variable x is one inspired by the method of least squares and called the "variance" or "dispersion" of x.

The variance of x is the expression $\sigma^2(x)=E(x-\alpha)^2$, where $\alpha=E(x)$.

(The square root of the variance is called the "standard deviation.") In words: take the square of the deviation of x from its expected value α, and use the sum (weighted sum, integral) of these squared deviations as a measure of clustering. Since a sum of squares vanishes only if each term does, the vanishing of the variance indicates that x is identically equal to its expected value (except perhaps for a set of probability zero). In general, the smaller the variance the closer the values of x lie to $E(x)$.

Such numbers as $E(x)$ and $\sigma^2(x)$ yield partial information about the distribution of the values of x. Complete information would mean an answer to every question of the form "what is the probability that x lies in the interval (α, β)?" In order to deal with such questions we introduce the notion of distribution function.

The distribution function $F_x(\lambda)$ of a random variable x is a function of a real variable λ defined for each λ to be the probability that $x<\lambda$.

These functions can be used to answer every probability question concerning random variables; for example the expression $F_x(\beta)-F_x(\alpha)$ represents the probability that x belong to the (half open) interval $\alpha\leqq x<\beta$, and the Stieltjes integrals $\int_{-\infty}^{\infty}\lambda dF_x(\lambda)$ and $\int_{-\infty}^{\infty}\{\lambda-E(x)\}^2dF_x(\lambda)$ represent the expectation and variance of x respectively. Distribution functions are useful because being comparatively simple real functions of real variables they are amenable to treatment by the methods of classical analysis. It is the whole purpose of a large part of probability theory to find the distribution functions of certain random variables.

10. Independent variables. Let us consider next two random variables x and y which are comparable in the sense that they are both represented by measurable functions on the same measure space Ω, so that $x=x(\omega)$ and $y=y(\omega)$. It is easy to see that the function $E(x)$, being defined by an integral, is homogeneous of degree 1 and additive, that is $E(\lambda x)=\lambda E(x)$ for every real constant λ and $E(x+y)=E(x)+E(y)$. Similarly the variance $\sigma^2(x)$ is homogeneous of degree 2, that is $\sigma^2(\lambda x)=\lambda^2\sigma^2(x)$. One way to prove this latter fact is to make use of the following identity connecting σ^2 and E:

$$\sigma^2(x) = E(x)^2 - E^2(x), \tag{1}$$

(where for later convenience we write $E(x)^2$ for $E(x^2)$ and $E^2(x)$ for $\{E(x)\}^2$). This identity in turn follows from the definition of σ^2. Since $\sigma^2(x)=E(x-\alpha)^2$

where $\alpha = E(x)$, we have $\sigma^2(x) = E(x^2 - 2\alpha x + \alpha^2) = E(x^2) - 2\alpha E(x) + \alpha^2$. (We used here the fact that the expected value of a constant is equal to that constant.) The identity (1) follows by substituting for α its value $E(x)$. Letting the formalism guide us we may inquire whether σ^2 is additive, that is whether or not the identity

$$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y) \tag{2}$$

is valid. The answer in general is no. In order to investigate conditions under which (2) is true we proceed to a brief discussion of some possible relations between pairs of random variables.

Let a and b be two independent events and let x and y be the associated counting random variables (so that x for example is 1 if and only if a occurs and $x = 0$ otherwise). The product random variable xy in this case can be equal to 1 if and only if both a and b occur, so that xy is the counting variable of $a \cap b$. Since $E(x) = P(a)$, $E(y) = P(b)$, and similarly $E(xy) = P(a \cap b)$, we have in this special case

$$E(xy) = E(x)E(y). \tag{3}$$

The validity of this formula is sufficiently important in the applications of probability to bear a name of its own: two random variables, not necessarily the counting variables of a pair of independent events, satisfying it are called "uncorrelated." The reason for the terminology is that the coefficient of correlation $r = r(x, y)$ of two random variables x and y is defined by $r = \{E(xy) - E(x)E(y)\}/\sigma^2(x)\sigma^2(y)$; this coefficient vanishes if and only if (3) holds.

It is now easy to state the facts concerning the formula (2): it is valid if and only if (3) is. In other words the variance is additive for a pair of random variables if and only if the expectation is multiplicative, that is if and only if they are uncorrelated. For the proof we merely expand the left member of (2), thus:

$$\begin{aligned}\sigma^2(x + y) &= E(x + y)^2 - E^2(x + y) \\ &= \{E(x)^2 - 2E(xy) + E(y)^2\} - \{E^2(x) - 2E(x)E(y) + E^2(y)\} \\ &= \sigma^2(x) + \sigma^2(y) - 2\{E(xy) - E(x)E(y)\}.\end{aligned}$$

Let us now return to the pair of counting variables x and y associated with two independent events a and b. Because of the independence of a and b, any probability statement concerning y is unaffected by our knowledge of ignorance of the value of x. More precisely, any two events defined by x and y, for example the events "$x = 0$" and "$y = 1$," are independent. If in general any two events by two random variables x and y respectively, that is any two events defined by inequalities of the form $\alpha \leqq x \leqq \beta$ and $\gamma \leqq y \leqq \delta$, are independent events, no matter what α, β, γ and δ are, we say that x and y are independent random variables. It is not too difficult to generalize what we proved about the special case of counting variables: for independent random variables the expectation, if it exists, is multiplicative and consequently the variance is additive. In still

other words: independence implies absence of correlation—a proposition which certainly sounds natural enough.

One word of caution before we leave this brief introduction to the notion of independence for random variables. What we defined was the independence of two random variables. It would be natural to try to define the independence of a finite or infinite sequence of random variables $x_1, x_2, \cdots$, by the requirement that any pair be independent. Natural, but as it happens, not very useful. The correct definition replaces two-term products by many-term products in the following way.

The random variables $x_1, x_2, \cdots$, are independent if the probability of the simultaneous occurrence of any finite number of the events defined by $\alpha_n \leqq x_n \leqq \beta_n$ is the product of the separate probabilities, no matter what real constants the α's and β's are.

It is easy to construct examples to show that this notion is indeed different from the notion of pairwise independence.

11. Law of large numbers. We are now in a position to reformulate and generalize the strong law of large numbers in terms of random variables. Let the sequence space of points $\omega = \{\omega_1, \omega_2, \cdots\}$ be the analytic model of the infinite repetition of an experiment one of whose possible outcomes is the event a. Let a_n be the event "ω_n belongs to a" or equivalently the event "the nth experiment results in a," and let $x_n = x_n(\omega)$ be the counting variable associated with a_n. In this context that means that $x_n(\omega)$ has the value 1 for all those sequences $\omega\{\omega_1, \omega_2, \cdots\}$ for which the nth coordinate ω_n belongs to a, and $x_n(\omega)$ has the value 0 otherwise. What significance has the sum $x_1 + \cdots + x_n$? Since a particular term x_j contributes one unit to this sum if and only if the jth experiment results in a, it is clear that the value of the sum, for any sequence ω, is the number of those coordinates among the first n coordinates of ω which do belong to a. But this is exactly the function we denoted above by $m_n(\omega)$. Hence our version of the law of large numbers is equivalent to the assertion that the averages $(x_1 + \cdots + x_n)/n$ converge (except possibly for a set of ω's of probability zero) to the constant $p = P(a)$. For the generalization of this result that we are about to formulate it is worth while to observe that $p = E(x_n)$ is also equal to the common value of the expectations of the x's.

The sequence of random variables $x_1, x_2, \cdots$ has two important properties which are sufficient to ensure the validity of the law of large numbers. One of these properties is independence. It follows very easily from the fact that the experiments yielding the values of the various x's are independently performed, that the variables $x_1, x_2, \cdots$ are indeed independent. The other essential property of the sequence is usually expressed by the statement that the random variables x_n all have the same distribution. The definition of this concept is as follows.

Two random variables x and y have the same distribution if for every interval (α, β) the probabilities of the two events $\alpha \leqq x \leqq \beta$ and $\alpha \leqq y \leqq \beta$ are equal, or equivalently if the distribution functions $F_x(\lambda)$ and $F_y(\lambda)$ are identical.

In our particular case it is the fact that the probability that ω_n belong to a is the same for all n (namely $P(a)$) that implies that the x_n all have the same distribution. That independence and equidistribution are indeed the crucial hypotheses for the law of large numbers is shown by the following general formulation of that law.

If $x_1, x_2, \cdots$ is a sequence of independent random variables with the same distribution, and if the expectations $E(x_n)$ exist and have the value α (necessarily the same for all n) then the averages $x_1+\cdots+x_n/n$ converge as $n\to\infty$ (except perhaps on a set of probability zero) to the constant α.

12. Central limit theorem. Sums (such as $x_1+\cdots+x_n$) of independent random variables with the same distribution occur very often in probability theory. It is of considerable practical importance to investigate the precise distribution of such sums and if possible the limiting behavior of these distributions. We assume concerning the x's that their expectations and variances both exist and write $E(x_j)=\alpha$, $\sigma^2(x_j)=\beta$. It follows from the independence and equidistribution of the x's that $E(x_1+\cdots+x_n)=n\alpha$ and $\sigma^2(x_1+\cdots+x_n)=n\beta$. At first sight this seems like a discouraging phenomenon: if both the expectation and the variance become infinite, how can we expect a reasonable asymptotic behavior from the much more delicate distribution function? But the way out of the difficulty is easy: by a translation and a change of scale (different to be sure for each n) it is possible to normalize the sum $x_1+\cdots+x_n$ so that its expectation is 0 and its variance 1 for every positive integer n. To get the expectation to be 0 we merely subtract its actual value, $n\alpha$, from the sum—the additivity of the expectation ensures the desired result. To get the variance to be 1 we divide by a constant factor. It is important to recall that the variance is homogeneous of degree 2, so that the constant factor will be not $n\beta$ but $\sqrt{n\beta}$. We arrive thus at the normalized sums

$$\frac{x_1+\cdots+x_n-n\alpha}{\sqrt{n\beta}}$$

and inquire again after the distribution function of this random variable and the limit of such distribution functions. The answer here is known and is embodied in the so called central limit theorem (or Laplace-Liapounoff theorem) stated as follows.

If $x_1, x_2, \cdots$ is a sequence of independent random variables with the same distribution, expectation α, and variance β, then the distribution functions of the modified sums $(x_1+\cdots+x_n-n\alpha)/\sqrt{n\beta}$ converge as $n\to\infty$ to a fixed distribution function, the same no matter what the original distribution of the x's is. In more detail, the limit as $n\to\infty$ of the probability of the event defined by the inequality

$$\frac{x_1+\cdots+x_n-n\alpha}{\sqrt{n\beta}}<\lambda$$

exists and is equal to

$$G(\lambda)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\lambda}e^{-u^2/2}du.$$

The distribution function $G(\lambda)$ is called the Gaussian or normal distribution.

With this statement we end our discussion of the development of probability theory and turn to a few remarks connected with the problem of application.

13. Determination of initial probabilities. When the mathematician announces that the probability of an event is a certain number, he is immediately faced with two questions. First the practical man asks what is the practical meaning of a probability statement? How should one act on it? If the mathematician succeeds in answering this question then the philosopher wants to know the reason for the answer. What establishes the connection between mathematical theory and practice? Our remarks in what follows will bear on these very old and very difficult questions only incidentally—they are dedicated mainly to a smaller problem of the theory, but one which frequently worries the layman.

The problem is how the probability of concretely given events is really defined. It is all very well to talk about Boolean algebras and measure theory, but what is the probability that a coin will fall heads up? What the layman realizes and what we now wish to emphasize is that the mathematician has not answered any such questions. He cannot. He can no more say that the probability of obtaining two heads in succession with a coin is $\frac{1}{4}$ than he can say that the volume of a cube is 8. The volume of a cube is given by a formula. If the hypotheses under which the formula applies are verified and if the variables entering into the formula are given specific values then the volume of a cube can be calculated. In exactly the same sense the mathematical theory of probability is a collection of formulae which enable us to calculate certain probabilities assuming that certain other ones are given. If we know that the probability of obtaining heads with a certain coin is $\frac{1}{2}$ and if we know that two successive tosses of the coin were performed independently then we can assert that the probability of getting two heads is $\frac{1}{4}$.

Despite the fact that probability theory shares with all other mathematical theories its inability to state a conclusion without hypotheses, the above answer to the layman's question will probably seem unsatisfactory to many readers. There must be some reason why most people believe that the probability of heads is $\frac{1}{2}$. It is often even proved. The usual proof is based on symmetry arguments, or equivalently on the principle of sufficient reason. (Why should heads have any greater likelihood of appearing than tails?) Do these proofs have any mathematical validity?

The answer is definitely yes. In some cases it is more pleasing to the intuition or more convenient for practice to formulate our hypotheses purely qualitatively. In almost all such cases the hypotheses take the form of invariance—the probabilities entering into the problem are required to be invariant under a certain group of transformations. It often turns out then that an existence and uniqueness theorem is true, that is it can be proved that there exists one and only one probability measure satisfying the stated hypotheses. Theorems of this type are certainly a part, an increasingly important part, of the theory of proba-

bility, and as long as their hypotheses are clearly formulated and recognized as hypotheses, the professional mathematician is the last person to sneer at them. Their advantage at the level of elementary pedagogy seems to lie in the fact that the statement "heads and tails are equally likely" is easier to grasp intuitively than the statement "the probability of heads is $\frac{1}{2}$."

We see thus that a mathematical statement on probability has to have certain either explicitly or implicitly given probabilities to begin with. In practice the physicist (or actuary, or anyone else interested in applying the theory) obtains these initial numbers experimentally. If he wants to know what is the probability of a coin falling heads up, he tosses the coin a large number of times and then uses the law of large numbers to assure himself that he may use the obtained frequency ratio as an approximation to the correct value of the probability. Or he may observe that the values of a random variable are obtained as the sum of a large number of independent variables each with a negligible variance and thus be led to introduce the normal distribution. Such approximative procedures are of course common to all parts of applied mathematics.

14. Conclusion. Our exposition is finished. If the reader has been patient enough to read this far he may be curious enough to read farther. Our scanty bibliography will furnish a basis for such reading. For certainly not all probability theory is contained in this paper, nor as yet in any collection of books or papers. There is still much room in the field for the exercise of the analytic ingenuity and abstract generality of both classical and modern mathematics. If this paper will be instrumental in persuading mathematicians that probability is mathematics, and in causing some to look into the subject more deeply than they had previously thought worth while, it will have more than accomplished its purpose.

BIBLIOGRAPHY

G. Birkhoff, Lattice Theory, New York, 1940.
H. Cramér, Random Variables and Probability Distributions, Cambridge, 1937.
A. Khintchine, Asymptotische Gesetze der Wahrscheinlichkeitsrechnung, Berlin, 1933.
A. Kolmogoroff, Grundbegriffe der Wahrscheinlichkeitsrechnung, Berlin, 1933.
S. Saks, Theory of the Integral, Warsaw, 1937.
J. V. Uspensky, Introduction to Mathematical Probability, New York, 1937

NON-ANALYTIC FUNCTIONS

SZU-HOA MIN, National Tsing Hua University

The triumph of the theory of analytic functions lies in the fact that it has wide applications not only in other branches of mathematics but also in many physical investigations. In regard to the latter, it is possible merely because many physical quantities are distributed like the values of a harmonic function,