PROBABILITY AS MEASURE

By J. L. Doob

University of Illinois

The following pages outline a treatment of probability suitable for statisticians and for mathematicians working in that field. No attempt will be made to develop a theory of probability which does not use numbers for probabilities. The theory will be developed in such a way that the classical proofs of probability theorems will need no change, although the reasoning used may have a sounder mathematical basis. It will be seen that this mathematical basis is highly technical, but that, as applied to simple problems, it becomes the set-up used by every statistician. The formal and empirical aspects of probability will be kept carefully separate. In this way, we hope to avoid the airy flights of fancy which distinguish many probability discussions and which are irrelevant to the problems actually encountered by either mathematician or statistician.

We shall identify as Problem I the problem of setting up a formal calculus to deal with (probability) numbers. Within this discipline, once set up, the only problems will be mathematical. The concepts involved will be ordinary mathematical ones, constantly used in other fields. The words "probability," "independent," etc. will be given mathematical meanings, where they are used.

We shall identify as Problem II the problem of finding a translation of the results of the formal calculus which makes them relevant to empirical practice. Using this translation, experiments may suggest new mathematical theorems. If so, the theorems must be stated in mathematical language, and their validity will be independent of the experiments which suggested them. (Of course, if a theorem, after translation into practical language, contradicts experience, the contradiction will mean that the probability calculus, or the translation, is inappropriate.)

The classical probability investigators did not separate Problems I and II carefully, thinking of probability numbers as numbers corresponding to events or to hypothetical truths, and always referring the numbers back to their physical counterparts. The measure approach to the probability calculus has put this approach into abstract form, and separated out the empirical elements, thus removing all aspects of Problem II. We shall explain this approach first in a simplified set-up, that which will be made to correspond (Problem II) to a repeated experiment in which the results of the nth trial can be any integer x_n between 1 and N (inclusive), in which the experiments are independent of each other, and performed under the same conditions. (The set-up will be applicable, for example, to the repeated throwing of a die.)

The measure approach treats this experiment as follows. Let $\omega:(x_1, x_2, \cdots)$ be any sequence of integers between 1 and N, inclusive. We consider ω as a point in an infinite dimensional space Ω. (Each point ω may be considered as a logically possible sequence of results of the given experiment, and this fact will guide us in solving Problem II.) A measure function is defined on certain sets of points of Ω as follows. Let $p_1, \cdots, p_N$ be any numbers satisfying the conditions

$$p_j \geqq 0, \qquad j \geqq 1, \qquad p_1 + \cdots + p_N = 1.$$

(How these numbers are chosen in any particular problem will be explained below. The method of choice is irrelevant to the mathematics, but is involved in the solution of Problem II.) The set of all sequences beginning with $x_1 = \alpha$ is given measure p_α. More generally, the measure of the set of all sequences beginning with $x_1 = \alpha_1, \cdots, x_n = \alpha_n$, is defined as $p_{\alpha_1} \cdot p_{\alpha_2} \cdots p_{\alpha_n}$. In this way, as can be shown,[1] a completely additive measure function is determined on certain point sets of Ω, on a field $\mathfrak{F}$ of sets so large that all the usual Lebesgue measure and integration theory is applicable. This means that there is a collection $\mathfrak{F}$ of sets of points of Ω such that if $S_1, S_2, \cdots$ are finitely or infinitely many sets in the collection, their sum $\sum_1^\infty S_n$, their intersection $\prod_1^\infty S_n$, and their complements are also in the collection. Each set S in $\mathfrak{F}$ has a definite measure $P(S)$, $0 \leqq P(S) \leqq 1$, and if $S_1, S_2, \cdots$ are finitely or infinitely many disjunct sets in $\mathfrak{F}$,

$$P(S_1 + S_2 + \cdots) = P(S_1) + P(S_2) + \cdots.$$

Problem II, the translation problem, is solved as follows. Each relevant event is made to correspond to a point set of Ω. A relevant event is a physical concept—defined by imposing some set C of conditions on the results of the experiments. The corresponding Ω-set is the set of sequences $(x_1, x_2, \cdots)$ satisfying the same set C of conditions, imposed on the x_j. Thus the set of all sequences beginning with $x_1 = \alpha_1$, $x_2 = \alpha_2$, is made to correspond to the event: *the result of the first experiment is* α_1, *of the second is* α_2. As is to be expected, the mathematical picture goes further than the real one. The "event" *1 occurs infinitely often in a sequence of trials* has only conceptual significance, physically, but the corresponding point set of Ω: the set of all sequences $(x_1, x_2, \cdots)$ containing infinitely many 1's, is a perfectly definite point set whose measure can be calculated in terms of $p_1, \cdots, p_N$. (In fact it is easily seen that this measure is 1 or 0, according as $p_1 > 0$ or $p_1 = 0$.) By "the probability of an event" we shall mean the measure of the corresponding Ω-set. As this measure has been defined, the probability that the nth trial results in a number j is p_j, and the probability that one trial results in j, and another in k, is $p_j \cdot p_k$.

[1] Cf. A. Kolmogoroff, *Ergebnisse der Mathematik*, Vol. 2, No. 3, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, where the most complete treatment of the approach to the probability calculus from the standpoint of measure is given.

The justification of the above correspondence between events and Ω-sets is that certain mathematical theorems can be proved, filling out a picture on the mathematical side which seems to be an approximation to reality, or rather an abstraction of reality, close enough to the real picture to be helpful in prescribing practical rules of statistical procedure. The following two theorems are important ones, from this point of view. These two theorems depend in no way on observed facts. They are stated and proved in the customary language of modern analysis.

THEOREM A: Let j_n be the number of the first n coordinates of the point $\omega:(x_1, x_2, \cdots)$ which are equal to j, where j is some integer $(1 \leqq j \leqq N)$ which will be kept fixed throughout the discussion. Then $0 \leqq j_n \leqq n$, and j_n varies from point to point on $\Omega: j_n = j_n(\omega)$ is a function of ω, that is of the sequence $(x_1, x_2, \cdots)$. When $n \to \infty$, j_n/n has not a unique limit independent of the sequence $(x_1, x_2, \cdots)$ under consideration. In fact if ω is the point $(k, k, \cdots)$, $j_n(\omega) = 0$ for all n, unless $j = k$; if ω is the point $(j, j, \cdots)$, $j_n(\omega) = n$ for all n. It is simple to give examples of sequences $\omega:(x_1, x_2, \cdots)$ for which $j_n(\omega)$ oscillates without approaching a limit, as $n \to \infty$. But Theorem A (usually called the strong law of large numbers) states that there is a set of sequences, i.e. an ω-set S, of measure 0, such that

$$\lim_{n\to\infty} \frac{j_n(\omega)}{n} = p_j, \tag{1}$$

unless ω is in S. In other words the sequences for which (1) is not true are exceptional in the sense of measure theory. If a new choice $\{p_i'\}$ of p_i's is made, then if $p_j' \neq p_j$, the new exceptional set includes all the sequences which were not exceptional before, since the limit in (1) becomes p_j'. Thus S depends essentially on p_j. Theorem A is a generalization of Bernoulli's classical theorem which states in our language that the measure of the set of sequences $\omega:(x_1, x_2, \cdots)$ for which

$$| j_n(\omega)/n - p_j | > \epsilon$$

approaches 0, as $n \to \infty$, for any positive ϵ. Theorem A is stronger because it states that there is actual convergence, whereas Bernoulli's theorem only concludes that there is a kind of convergence on the average.

Theorem A corresponds to certain observed facts, relating to the clustering of "success ratios," giving rise to empirical numbers $\bar{p}_j$. If the statistician wishes to apply his calculus to a given experiment (Problem II), he sets $p_i = \bar{p}_j$. There has been frequent discussion of the problem of determining the $\bar{p}_j$. This discussion of the $\bar{p}_j$ is sometimes held on so high a plane that the innocent bystander may wonder to what purpose such abstract philosophic concepts could possibly be put—besides that of stimulating further discussion on a still higher plane. The principle purpose of this paper is to discuss Problem I, but a few words on Problem II might not be out of place here. Almost everyone who is going to use probability numbers, the $\bar{p}_j$, *for other than conversational purposes,* derives them in the same way. There is a judicious mixture of experiments with reason founded on theory and experience. Thus if a coin is tossed by an experimenter who has examined the coin, and found that it had heads on one side but not on both, that it seemed balanced, and that (as a confirming check) tossing a hundred times gave around 50 heads, the experimenter would use $\frac{1}{2}$ as the probability of obtaining heads in his further reasoning. Of course there is no logic compelling this. The experimenter may have been fooled. A coin far out of balance may turn up 50 heads in 100 throws. But man must act, and the above procedure has been found useful, which is all that is desired. In many experiments, less reliance can be placed on a preliminary physical examination of the experimental conditions, and more must be placed on the actual working out of the experiment, as in the analysis of machine products. In that case, the actual results must be examined with great care, before attempting to use the above mathematical set-up. It sometimes may even be possible to change the experimental conditions to make the mathematics applicable.[2] In all cases, such mathematical theorems as Theorem A and the following Theorem B give the basis for applying the formal apparatus to practice. Indeed, the criterion of application includes the verification of special cases of the practical versions of Theorems A and B.

THEOREM B: Let $f_n(x_1, \cdots, x_{n-1})$ $(n > 1)$ be any function of the indicated variables, except that we suppose f_n only takes on the values 0, 1. Let $\omega: (x_1, x_2, \cdots)$ be a given point of Ω. Let n' be the number of the first n integers i such that $f_i(x_1, \cdots, x_{i-1}) = 1$, and let j'_n be the number of the first n integers i such that $f_i(x_1, \cdots, x_{i-1}) = 1$, and $x_i = j$. Then j'_n, n' are functions of $\omega: (x_1, x_2, \cdots)$. If $f_1 \equiv f_2 \equiv \cdots \equiv 1, j'_n = j_n, n' = n$, where j_n is as defined above. Suppose that there is an Ω-set S_0 of measure 0 such that $n' \to \infty$, as $n \to \infty$, unless $\omega \in S$. Theorem B states that there is then an Ω-set S' of measure 0, such that if $\omega: (x_1, x_2, \cdots)$ is not in S',

$$\lim_{n\to\infty} \frac{j'_n(\omega)}{n'} = p_j. \tag{1'}$$

(The set S' will depend on the given functions $f_1, f_2, \cdots$ and on the p_i, but is fixed, once these have been chosen.) This mathematical theorem corresponds to certain observed facts (usually summarized by stating that no (successful) system of play is possible). In fact, it states, in the language of practice, that rejecting certain trials, using as a criterion of acceptance or rejection the results of preceding trials, rejecting the ith trial if $f_i(x_1, \cdots, x_{i-1}) = 0$, does not affect the outcome of a game of chance, or, more precisely, does not affect the validity of the physical fact corresponding to Theorem A. If $f_1 \equiv f_2 \equiv \cdots \equiv 1$, (1′) becomes (1). The hypothesis that $n' \to \infty$ as $n \to \infty$ unless $\omega \in S_0$ is made to insure that infinitely many trials will be accepted. As an example of the

[2] Cf. W. A. Shewhart, *Statistical Method from the Viewpoint of Quality Control*, Washington, 1939.

possible variety in the definition of the f_i, we might define f_i as 1 if $x_{i-1} = N$, and $f_i = 0$ otherwise, so trials are accepted only if the previous trial resulted in the number N. Or much more complicated systems can easily be devised in which the criterion of acceptance of the nth trial depends on a varying number of the results of preceding trials. This theorem gives a mathematical counterpart to the physical idea of the mutual independence of repeated trials.

To summarize, mathematically (Problem I) the study has been reduced to that of the measure properties of Ω. This can be considered independently of any physical correspondence. The physical correspondence (Problem II) makes any event $\mathfrak{E}$ correspond to a point set E of Ω, the "probability of $\mathfrak{E}$" becomes the measure of E. Thus "the probability that the result of the first experiment is 3" becomes the measure of the set of sequences $(x_1, x_2, \cdots)$ beginning with $x_1 = 3$. *We have given no sharp definition of probability as a physical concept.* If the above mathematical set-up, after translation, using some set of p_j's, seems to fit a given physical set-up, any event will be said to have as its probability, the measure of the corresponding Ω-set. We have attempted to give no intrinsic a priori definition of the probability of an event: such a definition is quite unnecessary for our purposes. All that was required was a basis for prescribing the usual statistical procedures, and we have described such a basis.

In the above example, there would have been no new difficulty introduced if the x_n were not restricted to integral values, but allowed to take on any numerical values. The general point $\omega: (x_1, x_2, \cdots)$ of Ω would now be any sequence of real numbers. Instead of choosing the numbers $p_1, \cdots, p_N$ we choose a "distribution function" $F(x)$, a monotone function with the following properties:

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to +\infty} F(x) = 1, \qquad F(x - 0) = F(x).$$

Measure on Ω is defined as follows. The set of all sequences beginning with x_1 such that $a \leqq x_1 < b$ is given measure $F(b) - F(a)$. (The number $F(b)$ is called "the probability that $x_1 < b$.") More generally, the measure of the set of all sequences $(x_1, x_2, \cdots)$ beginning with $x_1, \cdots, x_n$, such that $a_j \leqq x_j < b_j, j = 1, \cdots, n$ is defined as $\prod_j [F(b_j) - F(a_j)]$. Thus if $F(x)$ defines a simple rectangular distribution: $F(x) = 0$ for $x < 0$, $F(x) = x$ for $0 \leqq x \leqq 1$, $F(x) = 1$ for $x > 1$, Ω-measure becomes (infinite dimensional) volume in the (infinite dimensional) unit cube. The correspondence (Problem II) between events and point sets of Ω is defined just as before. Sometimes it may be useful, in considering experiments giving rise to pairs of numbers, to let each x_n be a pair of numbers so that Ω becomes a sequence of points of a plane instead of a sequence of points of a line. In all cases there are mathematical theorems true of the resulting Ω which guide us (Problem II) in deciding just how the Ω-measure is to be defined, that is, how $F(x)$ is to be defined, in dealing with a given practical problem. But the essential point is this. Once Ω-measure has been defined, no changes or further hypotheses are possible or necessary. All relevant probability questions are answerable. Thus consider a question of the following type: if the experiments are grouped in some way,[3] with what probability will the groups have some given regularity property?[4] The question singles out a set E of sequences of Ω and asks: what is the measure of E? The problem may or may not be difficult mathematically,[5] depending on the grouping, but the original definition of measure on Ω needs no enlargement to answer it.

Technically, the mathematics has become the mathematics of a special type of measure defined on a space of infinitely many dimensions. If, however there is an integer ν such that only at most ν experiments are to be considered, we need only consider the ν-dimensional space of points $(x_1, \cdots, x_\nu)$, defining measure in this space in the same way as on Ω. Thus if x_n has the rectangular distribution defined above, the measure in $(x_1, \cdots, x_\nu)$-space becomes ordinary ν-dimensional volume in the unit cube. Perhaps the most common measure a statistician considers is that in which the measure of an $(x_1, \cdots, x_\nu)$-set E becomes "the probability that the point $(x_1, \cdots, x_\nu)$ representing an independent sample of ν from a normal distribution of mean 0 and variance σ^2" will lie in E:

$$(2) \qquad P\{E\} = \sigma^{-\nu}(2\pi)^{-\frac{1}{2}\nu} \int \cdots \int_E e^{-\frac{1}{2}(x_1^2+\cdots+x_\nu^2)/\sigma^2} \, dx_1 \cdots dx_\nu .$$

This example makes it obvious that the statistician is always doing measure theory, even though he may not state that fact explicitly. If the number of experiments has no upper bound conceptually—mathematically when the number of dimensions ν may increase without limit, as in Theorems A, B, it is much more convenient to use the space Ω, in terms of which experiments with varying numbers of trials can be considered simultaneously. The classical proofs of probability theorems, such as Bernoulli's theorem (the law of large numbers) are perfectly correct. If the "probability of an event" is interpreted as the measure of a set, these proofs do not even need verbal changes. There can be no question of the need for any axiomatic development beyond that necessary for measure theory, and the probability calculus can lead to no contradiction, unless the theory of measure is faulty.

It is customary for probability theorists to stop their discussions when the present stage is reached, so that the beginnings of a formal calculus have been constructed to deal with a repetition of independent experiments, conducted

[3] A grouping is necessary, for example, when two players are playing a game in which two out of three wins in the trials win a game. The trials are then grouped into successive groups of two or three, depending on how they come out.

[4] Continuing the preceding note, the question might be: will the ratio (games won by player α)/(games played) approach a limit with probability 1, that is, for all of the original sequences $\{x_n\}$ except possibly some forming a set of measure 0?

[5] The answer to the question of the preceding notes is simple. If p is the probability that player α wins a trial, the ratio in question approaches $p^3 + 3p^2(1 - p)$, the probability that α wins a game, with probability 1.

under the same conditions. Perhaps this is because of the following widely held syllogism: probability is something dealing with random events; random events are events having no influence on each other; therefore .... Unfortunately mathematicians and statisticians must deal with many problems involving dependent probabilities, whose solutions require the most delicate and careful applications of modern analysis. The rudimentary calculi which the outsiders find esthetically or philosophically pleasing are usually either insufferably awkward or completely insufficient for the needs of professionals. There is a strange situation, which one observer has facetiously described somewhat as follows: it is true with probability 1 that the technical workers in probability use the measure approach, but that the writers on "probability in general" descendants of Carlyle's professor, do not consider this approach worth much more than a passing remark.[6] The following pages outline how our previous treatment is generalized to deal with problems in which it is desirable to have the distribution of x_j vary with j (so that physically the experiments are no longer the same), and in which the x_j do not have to correspond to the results of independent experiments. Some attempt will also be made to show how the modern mathematical theory of real functions is applied to the probability calculus.

Let $x_j = x_j(\omega)$ be the jth coordinate of the point $\omega:(x_1, x_2, \cdots)$. Then as the sequence $\omega:(x_1, x_2, \cdots)$ varies, x_j does also: $x_j(\omega)$ is a function of ω. The functions $x_1(\omega)$, $x_2(\omega)$, $\cdots$ are functions defined on Ω, an abstract space on which a measure has been defined. Moreover Ω-measure has been defined in such a way that the Ω-set for which $x_j(\omega) < K$ (j, K fixed) is an Ω-set whose measure has been defined. (This set is composed of all sequences $(x_1, x_2, \cdots)$ whose jth coordinate is $<K$, and the measure is $F(K)$, using our last definition of Ω-measure.) In the terminology of measure theory, $x_j(\omega)$ is thus a measurable function. The study of the measure relations of Ω, and this is the whole of our probability calculus, can be considered, from this point of view, as the study of the properties of a sequence of measurable functions, one with very special properties, as we shall see, defined on some space. A measurable function defined on Ω is usually called a chance variable, in the theory of probability. (This terminology is somewhat dangerous, because it mixes Problems I and II.) The whole apparatus of modern real variable theory is applicable to these chance variables. Thus if $f(\omega)$ is a chance variable (measurable function of ω) (physically, a function of the observations), it is customary to define a number called its expectation. This number is simply the integral of $f(\omega)$, with respect to the given Ω-measure. The fact that the expectation of the sum of two chance variables is the sum of their expectations is simply the familiar theorem that the integral of the sum of two functions is the sum of their integrals. Let $S(j, K)$ be the Ω-set defined by the inequality $x_j < K$. Up to now we have supposed

[6] This analysis, like every other probability statement, is only an approximation to reality, but a fairly close one.

that the measure of $S(j, K)$ is independent of j, that is that the distribution of x_j is independent of j. We have also supposed that[7]

$$(3) \qquad P\{S(1, K_1) \cdots S(n, K_n)\} = P\{S(1, K_1)\} \cdots P\{S(n, K_n)\}$$

for any positive integer n, and numbers $K_1, \cdots, K_n$. That is, we have supposed that $x_1(\omega), x_2(\omega), \cdots$ are mutually independent chance variables.[8] In fact probability measure on Ω has been defined just to make the foregoing two facts true. Mutual independence is a very strong hypothesis to impose on a sequence of functions. In many probability problems (Markoff chains for example), more general measures must be defined on Ω. The sequence $x_1(\omega)$, $x_2(\omega), \cdots$ whose properties are those of Ω-measure, is then no longer a sequence of independent functions, and the distribution of x_j can vary with j.

At this level, the study becomes the study of any sequence of measurable functions, defined on some space of total measure 1. If f, g are given chance variables, they may turn out to be independent. In that case the theorem that the expectation of their product is the product of their expectations becomes, when translated into mathematical language, the familiar theorem that

$$\int\int f(x)g(y)\,dx\,dy = \int f(x)\,dx \int g(y)\,dy.$$

The mathematical theorems are not simply analogues of the probability theorems—they themselves are those theorems. When stated mathematically, the probability theorems need no proof: they need only recognition as standard results.

Empirical needs suggest that certain functions called conditional probability distributions, and conditional expectations, should be defined in a certain way. This is possible, as a formal matter,[9] and the theorems then proved about these functions gives them their usual meaning when translated into practical language. These functions are extremely useful tools in dealing with mutually dependent (that is not independent) chance variables.

The above approach is easily generalized to the stage needed in the study of Brownian movements or of time series, in which, instead of the proper initial

---

[7] $P\{S\}$ was defined as the measure of the Ω-set S.

[8] The n chance variables $f_1(\omega), f_2(\omega), \cdots, f_n(\omega)$ are said to be independent if for every set of n numbers $K_1, \cdots, K_n$, the following equality is true.

$$P\{f_j(\omega) < K_j, \quad j = 1, \cdots, n\} = \prod_j P\{f_j(\omega) < K_j\},$$

where $P\{\cdots\}$ denotes the Ω-measure of the Ω-set defined by the conditions in the braces. Thus in the example of a normal distribution in ν dimensions given above, $x_1, \cdots, x_\nu$ are independent functions on the space of ν dimensions, a fact which follows readily from the fact that the ν-dimensional density function is the product of ν functions of the separate variables.

[9] Cf. Kolmogoroff, loc. cit.

abstraction being a sequence $\{x_n\}$ of numbers, we have a one-parameter family $\{x_t\}$ (t takes on all real values). The number x_t may, for example, be thought of as the x-coordinate of a particle at time t. There is no difference in principle here: Ω is now the space of functions of t, instead of the space of sequences, that is functions of n. From the other point of view, instead of studying the properties of a sequence of measurable functions, it becomes necessary to study the properties of a one-parameter family of measurable functions.