CHAPTER I

FUNDAMENTALS

A. The Basic Assumptions (Sections 1-5)

1. *Introduction*

Probability calculus or probability theory is the mathematical theory of a specific area of phenomena, aggregate phenomena, or repetitive events. Certain classes of probability problems which deal with the analysis and interpretation of statistical inquiries are customarily designated as "theory of statistics" or "mathematical statistics."

Probability theory, as considered in this book, has nothing to do with questions such as: "Is there a probability of Great Britain some time in the near future being involved in a war with Egypt?" Similarly, a question concerning the probable historical truth of biblical narratives does not interest us. Nor are we concerned with the probability that the two poems called the Iliad and the Odyssey have the same author. Each of these questions, which can be discussed adequately from various points of view (political, sociological, historical, etc.—see also p. 3), deals with particular situations, and such questions, concerning "then and there," cannot be answered in our theory. Since the first task of any scientific endeavor is a limitation of its scope, we limit our scope, roughly speaking, to *a mathematical theory of repetitive events.*[1]

The probability concept used in probability theory has exactly the same structure as have the fundamental concepts in any field in which mathematical analysis is applied to describe and represent reality. Consider for example a concept such as velocity in mechanics. While velocity can be *measured* only as the quotient of a displacement s by a time t, where both s and t are finite, non-vanishing quantities, velocity in mechanics is *defined* as the limit of that ratio as $t \to 0$, or as the differential quotient ds/dt. It makes no sense to ask whether that differential quotient exists "in reality." The assumption of its mathematical existence is one of the fundamentals of the theory of motion; its justification must be found in the fact that it enables us to describe and predict

[1] We do not attempt to discuss in this book ideas and problems of "subjective probability."

essential features of observable motions. Likewise, no one has ever seen a straight line extend to infinity; but Euclidean geometry cannot be developed without the concept of a straight line extending to infinity in both directions.

In building a mathematical theory in any field of application, we start by stating certain basic assumptions, by means of which the fundamental concepts are introduced and defined. These concepts are idealizations (and in a way simplifications) of what is observable in nature. The mathematical analysis enables us to investigate the structure of these concepts and their interrelationships, thus deducing a system of propositions which form the mathematical theory. To make this construction scientifically meaningful, operational methods must be given that permit us to check the statements of the theory by means of observation.

2. *Sequences of Observations. The Label Space*

2.1. *Some basic concepts.* The subject of probability theory is long sequences of experiments or observations repeated very often and under a set of invariable conditions. We observe, for example, the outcome of the repeated tossing of a coin or of a pair of dice; we record the sex of newborn children in a population; we determine the successive coordinates of the points at which bullets strike a target in a series of shots aimed at a bull's-eye; or, to give a more general example, we note the varying outcomes which result from measuring "the same quantity" when "the same measuring procedure" is repeated many times. In every case we are concerned with a sequence of observations; we have determined the possible outcomes and recorded the actual outcome each time. In all these examples the outcome of each observation can be represented by a number, or by several numbers. If the original results are not numbers (as in coin-tossing and sex-recording), numbers may readily be substituted in various ways; thus "zero" and "one" may stand for "heads" and "tails," or for "male" and "female."

We call each single observation in such a sequence an *element* of the sequence, and the numerical mark that goes with it is called its *label value*, its *attribute*, or simply, the *result*. We assume that each time there are at least two possible results. The set of all conceivable outcomes of the observations is called the *label set*. Each outcome may be thought of as a "point" in a space known as the *label space* or *attribute space* (Merkmalraum[1]); many authors also use the term *sample space*. In the

[1] R. v. MISES, "Grundlagen der Wahrscheinlichkeitsrechnung," *Math. Z.* **5** (1919), pp. 52-99; p. 55.

game of tossing a coin, the label space may consist of the two points of the real line with the abscissas 0 and 1. In the throwing of a pair of dice (say a white die and a black die, where points on each die are observed separately) the label space may be considered two-dimensional and consisting of the 36 points (i, j), $i, j = 1, 2, 3, 4, 5, 6$. In the shooting of bullets onto a target, all points of the target form the two-dimensional label space. *Throughout Chapter I we will assume that the label set is finite or at most denumerable.* A denumerable or *denumerably infinite* or *countable* set is one whose elements can be put into a one-to-one correspondence with the sequence 1, 2, 3, ... of natural numbers.

In the abstract concept, which in our theory will take the place of the sequence of actual observations, only the *sequence of results*, of *labels*, will matter. We shall therefore take as the starting point the space S of possible label values and consider sequences of attributes or labels $\{x_j\}$, $j = 1, 2, 3, \ldots$, where each x_j is a point of S.

2.2. *Examples.* If S consists only of the two points 0 and 1, the sequence $\{x_j\}$ is an infinite sequence of zeros and ones, that is, for each j either $x_j = 0$ or $x_j = 1$. If the single experiment consists of the successive tossing of five coins and we consider as the result the number of heads in each experiment, then the label space consists of the six points 0, 1, 2, 3, 4, 5. If, however, in the same experiment we are interested in the precise succession of heads and tails for each group of five throws, then the label space consists of the 2^5 points (0, 0, 0, 0, 0), (0, 0, 0, 0, 1), ..., (1, 1, 1, 1, 1), and in the sequence $\{x_j\}$ each particular x_j is one of these 32 points.[2] If we toss a coin until heads appear for the first time and record the number of tosses required, the possible labels are the positive integers. Consider also the following tangible illustration of a countable label space. The surface of a cylinder is subdivided by lines parallel to its axis. The first part contains half of the surface, the adjacent one a quarter, the third adjacent one one-eighth of the surface, and so on indefinitely; the separation line belongs always to the smaller part and the remaining first line to the first part. The cylinder is rolled on a plane and we consider as a result the serial number of the part which touches the plane when the cylinder comes to rest. The labels are the positive integers.

Next consider the previously mentioned target. Let it be a circular disk divided by concentric circles into an inner circle A and four circular rings B, C, D, E. If we assume that each shot hits the target, the sample space consists of the labels A, B, C, D, E, or 1, 2, 3, 4, 5. Note that this

[2] If in this example we consider n rather than 5 tosses, the label space consists of 2^n points. If $n \to \infty$ the label space is no longer countable.

is not an example of a continuous two-dimensional label space: there are just five labels.

There is a difficulty characteristic of discrete label spaces, concerning the concept of "dimension." If we toss two coins a and b and observe for each of them whether the result is "heads" or "tails," it may seem natural to regard the label space as "two-dimensional," since there are *two chance observations*, one regarding the outcome of the first coin, the other that of the second coin. The four possible results are 0, 0; 0, 1; 1, 0; 1, 1. Each observation has two components and we are interested in each of the two components separately. In a similar way we may observe eye color and hair color; height and weight; height of father and height of son, etc. In each of these cases there are two variables, subject to chance, and one can observe one of them without observing the other. In such cases we speak of two-dimensional observations.

Of course, from a purely mathematical point of view, a set of discrete points can be considered of any dimension. In the example of the two coins (and similarly in each of the other examples), we may denote the possible outcomes 0, 0; 0, 1; 1, 0; 1, 1 by 1, 2, 3, 4, respectively (or in many other ways), and by means of such a vocabulary, which has nothing to do with the chance experiment under investigation, the four outcomes can be expressed in an unambiguous way. (An analogous remark applies, of course, to more than two dimensions.) Nevertheless, if we take into account the operational rule of the observation under consideration, it will in general be quite clear how to count the number r of observed chance variables and how to define an r-dimensional label set accordingly.

The following example will further illustrate our point. A die is rolled and we observe, in the usual way, the number appearing on the top face. Denote now by A a set consisting of the numbers 1, 2, 4, by B a set consisting of 2, 3, 4, 5. When the die is rolled one may then decide to ask whether "property A" holds or not and whether "property B" holds or not. If the die shows 2 or 4 both "properties" hold and this is denoted by AB; similarly, with A' for "not A," B' for "not B," the result 1 corresponds to AB', the results 3 and 5 to $A'B$ and 6 to $A'B'$. The definition of these four sets in terms of the results 1, 2, ..., 6—the vocabulary—*has nothing to do with the actual trial.* If we quote the result, which in the ordinary rolling of a die is one of the possible labels 1, 2, ..., 6, in terms of these pairs of sets, it is seen that we merely denote one-dimensional results by two-dimensional names. Denoting the result 6 by $A'B'$ does not make the observation "two-dimensional" (just as in the case of the two coins, the notation 2 for 0, 1 does not make this observation one-dimensional). In the same way we could define three or four or more

sets out of the six original results and classify a result as $ABCD$ or $A'BCD$, etc. The fact that a single point can be described as the intersection of r sets does not make the label r-dimensional.

3. Frequency. Chance

3.1. *First basic assumption.* Consider a finite label space S containing k distinct labels $a_1, a_2, \ldots, a_k$, and a sequence $\{x_j\}$, where each x_j is one of the a_i. Let a_i be a fixed label; among the first n elements of the sequence $\{x_j\}$ there will be a number n_i of results carrying the label a_i. The number n_i depends on n, on a_i, and on the sequence $\{x_j\}$. In our notation the subscript refers to the subscript of a_i and n to the total number of trials.[1] The ratio n_i/n is the *frequency* or relative frequency of a_i among the first n elements of $\{x_j\}$. For our purpose a frequency concept is needed that is independent of the number n. We therefore introduce our *first basic assumption. The sequence can be indefinitely extended and the frequency n_i/n approaches a limit as n approaches infinity.* We write

$$\lim_{n\to\infty} \frac{n_i}{n} = p_i, \qquad i = 1, 2, \ldots, k. \tag{1}$$

This limit, for which, necessarily, $0 \leqslant p_i \leqslant 1$, is then called the *limiting frequency* of the label a_i or also the *chance* of the label a_i within the sequence under consideration.[2] Similarly, in the example of the target (p. 3) we denote by $n_A, n_B, \ldots$ the number of shots out of the first n which hit the inner circle A, the first ring B, ...; then n_A/n, n_B/n, ... are the frequencies in the first n shots and $\lim_{n\to\infty} n_A/n = p_A$, etc.[3]

Note that our definition (1) shows the following analogy to that of velocity as the limit of the ratio s/t for t tending toward zero: while s/t depends on the duration t, the state of motion at one particular instant is characterized by $v = ds/dt$, which no longer depends on any duration.

We are conscious of the fact that a frequency need not reflect the magnitude of the limit. After $m = 1{,}000{,}000$ terms the relative frequency

[1] It would be more in keeping with the major part of the book to use N and N_i rather than n and n_i in this basic assumption, and to reserve n for the dimension of the collective (defined in Section 5) under discussion. If, for example, we consider the simultaneous tossing of n dice, the present situation should be described as N repeated tossings, each time of n dice. If we consider as labels the n numbers on the n upper faces, there are $6^n = k$ possible results. However, *this* n hardly appears in Chapter I and it would be inconvenient always to use N in this chapter.

[2] Similar ideas were presented by J. Venn, in his *Logic of Chance* (1866). See v. Mises [22], p. 22 ff. for more historical details.

[3] We shall also use the following notations for the limiting frequency: $p(a_i)$, $p(x_i)$, or $p(A)$, etc.

of "1" may, in principle, be near zero, although the limit may be *one*. It is a silent assumption *drawn from experience that in the domains where probability calculus has (so far) been successfully applied*, the sequences under consideration do not behave this way; they rather exhibit *rapid convergence*. This assumption does not follow from any rules or axioms of probability calculus. (See more on this point in Appendix Three, p. 108) A similar assumption is latent in all limit-definitions of applied mathematics.

Definition (1) remains unchanged if the label space S contains *countably many labels* and a_i is one of them. A sample space S which contains a finite or countable number of points is called *discrete*. For a discrete S our *first basic assumption* is now stated as follows:

Mathematical probability theory concerns infinite sequences $\{x_j\}$ *of labels in which each distinct label* a_i *has a limiting frequency* p_i; *this limit is also called the chance of* a_i *within the given sequence* $\{x_j\}$.

3.2. *Total chance.*[4] In the case of k labels $a_1, a_2, \ldots, a_k$, the following fundamental conclusion follows:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \lim_{n\to\infty} \frac{n_i}{n} = \lim_{n\to\infty} \sum_{i=1}^{k} \frac{n_i}{n} = \lim_{n\to\infty} \frac{n}{n} = 1. \tag{2}$$

We used the obvious relation

$$\sum_{i=1}^{k} n_i = n, \qquad \text{or} \qquad \sum_{i=1}^{k} \frac{n_i}{n} = 1. \tag{2'}$$

In the case of countably many labels we are still assured that the series $\sum_{i=1}^{\infty} p_i$ [each p_i defined by (1) and the summation over all points of S] converges. In fact,

$$\sum_{i=1}^{r} p_i = \sum_{i=1}^{r} \lim_{n\to\infty} \frac{n_i}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{r} n_i \leqslant \lim_{n\to\infty} \frac{n}{n} = 1,$$

so that the partial sums of $\sum_{i=1}^{\infty} p_i$ are seen to be non-decreasing and bounded above. However, $\sum_{i=1}^{\infty} p_i = 1$ does not follow from (1) (as can be shown by counterexamples). It needs an *additional assumption* or convention, as follows:

In the case of a discrete sample space

$$\sum_{i=1}^{\infty} p_i = 1, \tag{3}$$

[4] See footnote 1, p. 5.

where the sum is over all points in the sample space; in other words: the chance of all elements of S together equals one. Hence, always

$$p(S) = 1. \tag{3'}$$

This follows from (1) and (2) in the case of k labels, and is assumed in the case of denumerably many labels; in the latter case (3) is equivalent to the legitimacy of interchanging the limit $n \to \infty$ and the summation sign.[5]

It will be seen in Chapter II how our definition can be adapted to a non-discrete label space.

Problem 1. If all elements in a sequence have the same label A, the chance p_A equals 1. Is the converse also true?

Problem 2. If only two labels A and B exist, then $p_A + p_B = 1$. Is the converse also true?

4. *Randomness*

4.1. *General idea.* The notion of "chance" established so far leads to a value that is independent of the number of experiments. But the commonly accepted ideas of a game of chance imply another sort of independence also.

A boy repeatedly tossing a dime supplied by the U. S. mint is quite sure that his chance of throwing "heads" is $\frac{1}{2}$. But he knows more than that. He also knows that if, in tossing the coin under normal circumstances, he disregards the second, fourth, sixth, ..., turns, his chance of "heads" among the outcomes of the remaining turns is still $\frac{1}{2}$. He knows—or else he will soon learn—that in playing with his friends, he cannot improve his chance by selecting the turns in which he participates. His chance of "heads" remains unaltered even if he bets on "heads" only after "tails" has shown up three times in a row, etc. This particular feature of the sequence of experiments appearing here and in similar examples is called *randomness*. A more precise formulation will be given presently.

Consider the following infinite sequence consisting of alternating runs of zeros and ones of increasing length

$$01 \quad 0011 \quad 000111 \quad 00001111 \cdots \tag{4}$$

[5] The consideration of countably many labels has been emphasized by E. BOREL, *Rend. Circ. Mat. Palermo* **27** (1909), p. 247.

It is obvious that here the limiting frequency of the label one is $\frac{1}{2}$, but it would be very easy to increase this chance by a suitable selection: if only those turns that follow a turn which has resulted in "one" are considered (or, in other words, if a player bets only on trials following a "one"), then the frequency of the label "one" in the sequence of the elements thus selected will obviously approach unity. If, in sequence (4), pairs of two successive figures are considered (the labels now being 00, 01, 10, 11), the chance of getting 11 is by no means $\frac{1}{4}$ as in a heads or tails game, but $\frac{1}{2}$, and the chance of 01 is zero.

We give a second example. Consider the repeated tossing of a true coin, that is, one for which the two chances are equal. This leads to a sequence $\{x_j\}$ of zeros and ones as, for example:

$$0\,1\,0\,0\,0\,1\,1\,0\,1\,0\,0\,1\,1\,1\,0\,1\,0\,0\,0\cdots. \tag{a}$$

From (a) we derive a new sequence (b) by adding the first and second results, the second and the third, the third and the fourth, and so on. The resulting sequence is

$$1\,1\,0\,0\,1\,2\,1\,1\,1\,0\,1\,2\,2\,1\,1\,1\,0\,0\cdots. \tag{b}$$

In sequence (b) the possible labels are 0, 1, 2 and it can be shown that they appear with chances $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, respectively; but in (b) a "0" can never be followed by a "2" and a "2" never by a "0." In fact, consider three successive elements of (a): $x_{\nu-1}$, x_ν, $x_{\nu+1}$; if $x_{\nu-1} + x_\nu = 0$, then $x_\nu = 0$, and then $x_\nu + x_{\nu+1}$ cannot be "2." If, in contrast to (b), we consider the sequence originated by the repeated simultaneous throwing of two true coins with the number of heads as labels, the possible labels are again 0, 1, 2 and the respective chances are again $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$. But in this sequence the chance of the result "2" remains $\frac{1}{4}$ even if the gambler who bets on "2" decides to participate only in those games which follow a "0." In sequence (b) the chance of "2," which equals $\frac{1}{4}$ in the complete sequence, drops to zero by such a selection.

Actually, the destruction of randomness in sequence (b) is not limited to the impossibility of a succession 0, 2 or 2, 0. In (b) the succession 0, 1 is possible and likewise the succession 1, 0. But the three elements 0, 1, 0 cannot arise in succession since from $x_{\nu-2} + x_{\nu-1} = 0$, it follows that $x_{\nu-1} = 0$; then, necessarily, $x_\nu = 1$ and $x_\nu + x_{\nu+1}$ cannot be 0. In the same way, there are impossible sequences of h labels even if the subsequences of $h - 1$ labels are possible. An example for $h = 4$ is the sequence 0, 1, 1, 2, for $h = 5$: 0, 1, 1, 1, 0, etc.

The difference between sequences such as (4) and (b), which are both "sensitive to place selections," or which are not "random sequences"

is that (4) is a sequence constructed according to a mathematical rule, while (b) originated from the random sequence (a) by means of an operation which destroyed the randomness.[1]

4.2. *Place selection.*[2] In order to arrive at a more precise formulation of randomness we introduce the concept of *place selection.* This notion applies to a specific manner in which an infinite subsequence is derived from an infinite sequence of elements. We say that this *subsequence has been derived by a place selection if the decision to retain or reject the* nth *element of the original sequence depends on the number* n *and on the label values* $x_1, x_2, \ldots, x_{n-1}$, *of the* $(n-1)$ *preceding elements, and not on the label value of the* nth *element or any following element.* In a general way, a place selection can be defined by a set of functions

$$f_1, \quad f_2(x_1), \quad f_3(x_1, x_2), \quad f_4(x_1, x_2, x_3), \ldots, \tag{5}$$

each function taking on the value one or zero, where $f_n = 1$ means retaining the nth element and $f_n = 0$ means rejecting it. The rules must be such that the selected subsequence is infinite.[3] If the label values are a and b, the following is an example of a place selection: the first element x_1 is retained; x_i for $i > 1$ is retained if, and only if, i is a prime number and $x_{i-1} = a$. We then have

$$f_1 = 1; \quad f_2(a) = 1, \quad f_2(b) = 0; \quad f_3(x_1, a) = 1, \quad f_3(x_1, b) = 0;$$

$$f_4 = 0; \quad f_5(x_1, x_2, x_3, a) = 1, \quad f_5(x_1, x_2, x_3, b) = 0; \quad f_6 = 0; \quad \text{etc.}$$

An example of a place selection where the selection of the nth element depends upon all n trials is as follows. The labels are 0 and 1; an element x_n is selected if n is a complete square and if there are an even number of ones among the first $n-1$ results $x_1, x_2, \ldots, x_{n-1}$. A further example of a selection (where the f_n in (5) do not depend on the $x_1, \ldots, x_{n-1}$) is as follows: consider the binary expansion of a number α in (0, 1), where $\alpha = \alpha_1\alpha_2\alpha_3 \ldots$ is not a binary rational. In the sequence $\{x_j\}$ retain x_n if $\alpha_n = 1$, otherwise reject it.

We see that always a place selection defines a sequence of natural numbers $\alpha_1 < \alpha_2 < \ldots$ for which in (5) $f_{\alpha_\nu}(x_1, x_2, \ldots, x_{\alpha_\nu - 1}) = 1$.

From experience with sequences representing games of chance we

[1] A sequence like (b) is intermediate between a random sequence and a regular sequence, ust as "dependence" (see Sections 9 and 10 of this chapter, etc.) is in between "independence" and full (=analytic) regularity.

[2] Remember footnote 1, p. 5, here and throughout this chapter.

[3] For further discussion, see Appendix One. It is, for example, assumed that the complete set $\{f_n\}$ can be described in a finite number of words.

gather that in this type of sequences, the chance of a label is *insensitive to place selections*. For example, if we sit down at the roulette table in Monte Carlo (see Chapter IV, p. 155 for a description of roulette) and bet on red only if the ordinal number of the game is, say, the square of a prime number, the chance of winning (that is, the chance of the label red) is the same as in the complete sequence of all games. And if we bet on zero only if numbers different from zero have shown up fifteen times in succession, the chance of the label zero will remain unchanged in this subsequence. Insensitivity to place selection, it is seen, is equivalent to what may be called the *principle of impossibility of a successful gambling system*.

The banker at the roulette acts on this assumption of *randomness* and he is successful. The gambler who thinks he can devise a system to improve his chances meets with disappointment. Insurance premiums are successfully computed from formulas based on the assumption of randomness, and in a great number of statistical investigations (see, for example, the observations of "runs" in birth records, Chapter IV, Section 7), the theory based on this assumption proves to hold. We also add the following consideration. The "probability of casting 3" with a given die under given circumstances may be considered as a "physical constant" p of the die just as its density, etc. To find experimentally an approximate value of this p we cast the die repeatedly and note the results. Our randomness principle states that the chosen mode of notation of the results does not influence the value of p.

Thus the idea of our *second basic assumption* leads to the following statement:

In probability theory we shall deal (not exclusively) with such particular sequences satisfying the first basic assumption for which it may be assumed that the limiting frequency, or chance of any label a_i is insensitive to place selections. If this holds true, the chance will also be called probability, or more completely, the probability $p(a_i)$ of encountering the label a_i in the sequence under consideration.

The term "insensitivity to place selections" or *randomness* needs some further elaboration. If this is understood to mean "insensitivity to all conceivable place selections," then certain difficulties can be raised in connection with the use of the word "all."[4] These can, however, be

[4] A sequence of natural numbers, in increasing order, defines a place selection. Among "all" place selections, there are the place selections G consisting of the system of all sequences of positive integers in increasing order. If in the particular sequence K under consideration, the labels are 0 and 1 with $p(1)$ neither zero nor one, then there are infinitely many "ones" in K and the sequence of integers corresponding to the particular order of the ones in K is contained in G; for *this* selection, the limiting frequency of one is not $p(1)$ but unity. Therefore K cannot be insensitive to "all" place selections.

avoided: a place selection (5) is defined by a given mathematical law; the concept of a mathematical law has a clear meaning in the framework of a formalized logic within which there are not more than countably many mathematical laws,[5] hence not more than countably many place selections, and we require insensitivity only with respect to such place selections. *A place selection must state in an unambiguous way, for each* $\nu = 1, 2, \ldots$, *whether or not* x_ν *is to be retained.*

Actually, it is not contended that any observable sequence can be identified as one satisfying our assumptions. Rather certain conclusions derived from the basic notions and definitions (sometimes in a complicated way) find their counterparts in observable facts. This will be shown in detail in many examples throughout this book. The student familiar with the role of theoretical physics in any field of application will realize the complete analogy.

Problem 3. Prove that in sequence (4), given above, the frequency limit of "ones" exists and equals $\frac{1}{2}$. Prove that if the successive pairs (1st and 2nd, 3rd and 4th elements, etc.) are considered as new elements with the four different labels 00, 01, 10, 11, the frequency limits of 00 and 11 are each $\frac{1}{2}$ and the limits for 01, 10 are both zero.

Problem 4. Consider[6] the sequence $0^1\ 1^1\ 0^2\ 1^2\ 0^4\ 1^4\ 0^8\ 1^8 \cdots$. Show

(a) that the relative frequency of "one" equals $\frac{1}{2}$ for

$$n_\nu = 2[1 + 2 + 4 + \cdots + 2^{\nu-1}] = 2(2^\nu - 1), \quad \nu = 1, 2, 3, \ldots,$$

(b) that for $\bar{n}_\nu = n_\nu + 2^\nu = 3 \cdot 2^\nu - 2$ the relative frequency of "one" has the value

$$\frac{1}{2}\frac{n_\nu}{n_\nu + 2^\nu} = \frac{2^\nu - 1}{3.2^\nu - 2} < \frac{1}{3},$$

(c) that no limit of relative frequency exists for this sequence (and that any number between $\frac{1}{2}$ and $\frac{1}{3}$ is a limit point of relative frequencies).

5. *The Collective*

5.1. *Definition.* It is useful to denote a well-defined concept of the theory by a concise term. We shall apply the term *collective* to a long sequence of identical observations or experiments, each experiment leading

[5] The system G of the preceding footnote contains non-countably many place selections.

[6] By 0^2 or 1^4 we denote the sequences 00 or 1111, respectively.

to a definite numerical result, provided that the sequence of results fulfills the two conditions: existence of limiting frequencies and randomness. (Often the word *population* is used in a similar sense, although sometimes one refers to a population when it is left undecided whether randomness prevails.) The word collective will also be applied to the infinite sequence of numbers (labels) which forms the theoretical counterpart of the long sequence of experiments. As in other branches of science, we use the terms (collective, label, frequency, probability, etc.) on the one hand in their abstract mathematical sense, and on the other hand with respect to a concrete field of application, for example a game of chance, an insurance problem, a problem of genetics, or of theoretical physics.

Remembering the definition of randomness in terms of place selection (see Section 4.2) we now state the definition of a *collective* as follows.

Let $S = \{a_i\}$ be a discrete label set and $K = \{x_j\}$ a sequence of elements of S. Let G be a system of denumerably many place selections. We assume that

(1) *For every label a_i the limiting frequency p_i exists in K;*

(2) $\sum_{i=1}^{\infty} p_i = 1$, *the sum extended over all elements of S;*

(3) *Any place selection Γ belonging to G applied to K produces an infinite subsequence of K in which again, for every a_i, the limiting frequency exists and is equal to p_i.*

Then, K, or, more explicitly, $K(G, S)$, is called a collective (with respect to G and S) and $p(a_i) \equiv p_i$ is the probability of encountering the label a_i in K.

We obtain a concrete idea of the set G of place selections which are supposed not to change the frequency limits if we visualize G, for example, as follows: in G are contained all those place selections which present themselves in a particular problem under consideration. Obviously, not more than denumerably many place selection arise in any concrete problem.

We shall also assume explicitly that the set G has a certain "closure" property. If Γ_1 denotes a selection belonging to G and Γ_2 another selection of G, then the selection obtained by applying Γ_2 to Γ_1 or Γ_1 once more to Γ_1, etc., also belongs to G.

If we identify G with the totality of all place selections in a formalized logic R (see p. 11), then G will contain all place selections intervening in a particular problem A since any selection appearing in A will be defined in terms of the mathematical laws of R. For further discussion see Appendix One.

5.2. *Comments.* The concept of a collective is thus at the basis of our probability theory. The term probability is meaningful for us only with

regard to a clearly defined collective (or population). This is the meaning of our original assertion, Section 1, which says, in the present terminology, that any probability statement relates to a collective, an aggregate phenomenon, rather than to an individual isolated event. The probability of dying within the coming year may be p_1 for a certain individual if he is considered as a man of age 42, and it may be $p_2 \neq p_1$ if he is considered as a man between 40 and 45, and p_3 if he is considered as belonging to the class of all men and women in the United States.[1] In each case, the probability value is attached to the appropriate group of people, rather than to the individual. Another example: if one draws thirteen times from a box containing the 26 letters of the Roman alphabet and the letters obtained in succession form the word "Massachusetts," one would be inclined to say that something very surprising has happened, that the probability for this event to happen is extremely small. On the other hand, each of the $26^{13} \sim 2.4 \times 10^{18}$ combinations of 13 letters out of 26 has, under normal conditions of drawing, the same probability. The obvious answer to this apparent discrepancy is that in our first reaction we think of a collective with a two-point label set only, the labels being: meaningful word or meaningless combination of letters; the probability of the first label is then indeed very much smaller than that of the latter. If, however, the label set consists of all the 26^{13} possible combinations, all these labels have equal probability.[2]

A similar example: it is not true that in a lottery with 1,000,000 lots the result "900,000" is less probable than any other number, if the collective has as labels the number of the lot drawn; it *is*, however, an "improbable" result if the attribute, in a different collective, consists of the number ν of zeros, $\nu = 0, 1, 2, \ldots, 6$, at the end of the number of the lot drawn (here $\nu = 5$).

Colloquially, the word probability is used in a much wider sense (see p. 1). One speaks, for example, of the probability of making a train in certain circumstances. This "probability" can perhaps be given a numerical value if the circumstances enable one to establish a sequence of observable cases which, at least, in principle can be extended indefinitely. But, if one talks of the probability that the two poems known as the Iliad

[1] We apply here our theory to a large but finite population.

[2] Sometimes one reads that an event whose probability is 10^{-18} is "practically impossible." If this way of talking is consistently used, one would have to say that no combination of 13 letters is "practically possible." Likewise, *any* sequence of 100 trials with a (correct) die has the probability $(\frac{1}{6})^{100} \sim 10^{-88}$. Thence, by the above recipe no such sequence could appear. Of course, the same fallacy is contained in the statement: "An event whose probability is $1 - \epsilon$ (ϵ a very small number) is practically certain to occur."

and the Odyssey have the same author, no reference to a prolonged sequence of cases is possible and it hardly makes sense to assign a *numerical* value to such a conjecture.

Finally, consider the following distinction which, in one form or another, arises in any theoretical science. It is not the task of the theory of probability to determine the numerical values of the probabilities of certain events. Such probabilities play a role similar to that of the initial data (initial coordinates of a point; initial velocity, etc.) in mechanics; they are "given" and how we arrive at them does not concern the theory proper. They may have been obtained by experiment. Often they are the result of a combination of induction (a general knowledge of, say, the properties of a die) and a direct statistical experiment for fairly large n.[3] The contents of insurance tables form "initial data" in this sense and probability calculus uses these data in drawing mathematical conclusions. In some books there are discussions about how we know that the probabilities of all the faces of a correct die are equal. Such arguments are futile from our point of view just as it is a futile question for a student who wishes to compute the length of a diagonal in a regular hexagon to ask how we know that the six sides of the hexagon are equal. Likewise, geometry does not answer the question about the distance between two well-defined points on the surface of the earth; it shows only how this distance can be computed if other related magnitudes, distances, angles, etc., are known. In the same way, probability calculus establishes relations between probabilities in connected collectives.

Note that in any problem of probability calculus, the given initial data as well as the final data are probabilities. By certain operations to be discussed presently, new collectives are derived from given ones and if the probabilities in the original collectives are considered as known, one may ask for the probabilities in the derived collectives. *Probability calculus teaches us to compute the probability distributions in derived collectives from given distributions in the collectives from which they have been derived.*[4] The operations by means of which the new collectives are derived from given ones will be discussed in the remainder of this chapter.

Problem 5. Try to find reasonable definitions (by constructing appropriate collectives) for the following probabilities:

---

[3] In many cases the more important conclusions do not depend on the numerical values of the "given" probabilities.

[4] We shall see (Chapter I, Section 7.3, and examples in Chapter IV, etc.) that higher-dimensional collectives include dependent events. Hence the present statement is fairly general.

(a) probability of a newborn boy reaching the age of 70;

(b) probability of an American girl of 20 marrying.

B. The Operations (Sections 6-10)

6. *First Operation: Place Selection*

In the following sections of this chapter four simple types of operations on collectives (or transformations of collectives) will be discussed. By applying a finite or infinite number of those operations, we obtain again collectives and we submit that no other types of operations are necessary for the transition from collective to collective, a process by means of which most of the known problems of probability calculus are solved.[1]

The first of the basic operations is called *selection*. Let x_1 , x_2 , x_3 , ... be the successsive label values of a given collective, $K(G, S)$. The label set may consist of the points a_1 , a_2 , a_3 , ... so that each x equals one of the a's. The *probability distribution* consists of the numbers p_1 , p_2 , p_3 , ..., where p_i is the probability of a_i , and $p_1 + p_2 + p_3 + \cdots = 1$. Now let Γ be a place selection (Section 4.2) singling out the ordinal numbers α_1 , α_2 , α_3 , ... Consider the sequence K' of label values x_{α_1} , x_{α_2} , x_{α_3} , ... This new sequence $K' = \Gamma(K)$ is again a collective since (i) according to the assumption of randomness, the frequency limit for every label exists in K' and has the same value as in K, and (ii) a place selection Γ' operated on the new sequence K' amounts to a place selection $\Gamma'\Gamma$ on the original collective. The composition $\Gamma'\Gamma$ of the two place selections Γ and Γ', applied in this order, is expressed explicitly as follows. If Γ singles out the ordinal numbers α_1 , α_2 , ... in K and Γ' the numbers β_1 , β_2 , ... in K', then $\Gamma'\Gamma$ is a place selection α_{β_1} , α_{β_2} , ... operating on K. Hence $\Gamma'\Gamma(K)$ has the same distribution as $\Gamma(K)$ and as K. Thus, a new collective has been derived from the original one: the new sequence is a partial sequence of the original one, the label value of each element has been left unchanged, and the new distribution is seen to be the same as the one given. We used the "closure" property mentioned at the end of Section 5.1.

We make use of this transformation in almost every probability problem. Consider, for example, two people A and B casting alternately

[1] In Section 4.1, p. 8 we discussed an example of a non-collective derived from a collective, but, of course, *not* by one of the four operations we are going to discuss. The study of such non-collectives is by no means excluded in our theory.

the same dice. We assume not only that the total of all casts forms a collective, but also that the casts of A as well as of B are collectives with the same probability values. If the entries in a volume listing birth records, with specification of the sex of newborn children (labels 0 and 1), are considered as a collective (whether or not this is justifiable is a question of fact and of judgment), then we also assume that the entries, say, at the tops of the pages, taken by themselves, form a collective with the same distribution.

These and previous examples show that the introduction of the concept of randomness, in one way or another, is unavoidable. Some authors do not include insensitivity to place selections among the general features of probability sequences. Then they have to assume it in most individual problems. In substance, this procedure does not differ from ours. The principle of randomness as stated in Section 4 is equivalent to the general pronouncement that any place selection occuring in a problem will be assumed not to change the probabilities.

Consideration of sequences not obeying the randomness axiom will not be excluded but they will not be called collectives and the term probability will not be used for them. We have given examples in Section 4.1, in particular the sequence (b) discussed there in detail. The limiting frequency of an attribute is then called a chance. This distinction will be made throughout the book. (See, e.g., the end of Section 7.1 and the beginning of Section 8.2 of this chapter.)

We summarize the content of this section in the statement:

If from a collective $K(G, S)$ with probability distribution $\{p_i\}$ a subsequence K' is derived by a place selection of G, then K' is also a collective $K'(G, S)$ with the same probability distribution $\{p_i\}$.

Problem 6. Consider the fifth decimal place of the logarithms of the integers 1 to 500 taken from a seven-place table. Compute the frequency of zeros occuring at that place and compare it with the frequency of zeros in the following selected sequences:

(a) the logarithms of even integers;

(b) the logarithms of integers following a prime number;

(c) the logarithms whose fourth decimal place is one of the figures 1, 3, or 5.

7. Second Operation: Mixing. Probability as Measure

7.1. *Mixing finite numbers of labels.* Consider a collective $K_0(G, S)$ consisting of the sequence of observed label values $x_1, x_2, x_3, \ldots$ taken from the label set S. We now retain all elements, but change their label

values (replacing, say, the labels a_1, a_2, a_3 by b_1; a_4, a_5 by b_2; and all others by b_3).

More generally, let A be a subset of the label set S consisting of $k > 1$ elements of S. For ease of argument we shall suppose that A consists of the first r elements $a_1, a_2, \ldots, a_r$ of S. From K_0 we form a new sequence K by replacing those elements of K_0 which belong to A by the symbol b, while leaving unchanged the remaining elements of K_0. If among the first n elements of K_0 the label values $a_1, a_2, \ldots, a_r$ appear $n_1, n_2, \ldots, n_r$ times, respectively, the limits as $n \to \infty$ of n_1/n, $n_2/n, \ldots, n_r/n$ are the probabilities $p_1, p_2, \ldots, p_r$ of the $a_1, a_2, \ldots, a_r$ in K_0. In the derived sequence K the number n_b of elements with the new label b is clearly $n_b = n_1 + n_2 + \cdots + n_r$. Hence, the limit

$$\lim_{n\to\infty} \frac{n_b}{n} = p_b \tag{6}$$

exists and its value is

$$\lim_{n\to\infty} \frac{n_1}{n} + \lim_{n\to\infty} \frac{n_2}{n} + \cdots + \lim_{n\to\infty} \frac{n_r}{n} = p_1 + p_2 + \cdots + p_r . \tag{7}$$

Instead of n_b and p_b we may also write $n(A)$ and $p(A)$ in order to indicate that the label consists of the set A. Our result is then

$$p(A) = p_1 + p_2 + \cdots + p_r . \tag{8}$$

Thus the chance $p(A)$ exists in K and is equal to $p_1 + p_2 + \cdots + p_r$.

Moreover, K also fulfills the condition of randomness. In fact, were it possible to change the value of $p(A)$ by applying a place selection to K, the same place selection applied to K_0 would necessarily change at least one of the values $p_1, p_2, \ldots, p_r$. But this is impossible since K_0 is a collective. Of course, the probability of each label of S which is not in A is the same in K as in K_0. The same procedure by which the labels contained in A were fused may be applied to a second subset A' of $S - A$, etc.

Our reasoning shows so far that by fusing or "mixing" a finite number of label values in a collective, a new collective is formed. The distribution in the new collective is found by adding the original probabilities of all those label values which had been replaced by a single one

$$p(A) = \sum_{a_i \in A} p_i, \qquad p(A') = \sum_{a_j \in A'} p_j , \ldots, \tag{8'}$$

In more colloquial language, the mixing rule is known as the law of *either-or probability*: the probability of obtaining either a_1 or a_2 or a_3

(where the a_i are mutually exclusive) is the sum of the single probabilities for a_1, a_2, a_3. Consider, however, the following example. A tennis player knows from experience that he may have an 80% chance of winning a certain tournament in New York, and that his chance of winning another tournament held in London on the same day may be 70%. The events are mutually exclusive, but the chance in this "either-or" case is by no means $0.80 + 0.70 = 1.50$. One sees that the mixing rule cannot be stated in a strict way without using the notion of a collective. Only probabilities in one and the same collective may be added.

The scope of the mixing rule is wider than would appear from the preceding. First, it can be stated independently of any assumption about randomness. If it is only known that $p_1, p_2, \ldots, p_r$ are the chances for $a_1, a_2, \ldots, a_r$, the *chance* of the new label b is $p_1 + p_2 + \cdots + p_r$. Neither is the restriction to the fusing of a *finite* number of disjoint label values essential.

7.2. *General case.* We consider now *the problem of mixing an infinite number of labels*; each label may be denoted by a positive integer. Let A be a subset of S, consisting of any given set of positive integers. We shall prove that, in generalization of (8'),

$$\lim_{n\to\infty} \sum_{i\in A} \frac{n_i}{n} = \sum_{i\in A} p_i , \tag{9}$$

that is, the limit of the (relative) frequency of the new label (consisting of the integers in A) exists and equals the sum of the p_i for $i \in A$.[1]

Denote by A' the *complementary* set[2] to A, by A_k the set of those elements of A which are $\leqslant k$ and by A_k' the set of those elements of A' which are $\leqslant k$. Then

$$\sum_{i\in A} n_i + \sum_{i\in A'} n_i = n \tag{10}$$

and using (3)

$$\sum_{i\in A} p_i + \sum_{i\in A'} p_i = 1. \tag{11}$$

Also:

$$\sum_{i\in A} n_i \geqslant \sum_{i\in A_k} n_i , \qquad \sum_{i\in A'} n_i \geqslant \sum_{i\in A_k'} n_i . \tag{12}$$

[1] The following proof is due to G. Pólya (private communication).

[2] By $A' = S - A$ we mean the set of all elements of S not contained in A.

Observe that by Eq. (7) the assertion (9) is true if A is a finite set. From (12)

$$\varliminf_{n\to\infty} \frac{1}{n}\sum_{i\in A} n_i \geqslant \lim_{n\to\infty} \frac{1}{n}\sum_{i\in A_k} n_i = \sum_{i\in A_k} p_i$$

and this being true for all values of k

$$\varliminf_{n\to\infty} \frac{1}{n}\sum_{i\in A} n_i \geqslant \sum_{i\in A} p_i \,. \tag{13}$$

Likewise from (11) and (12),

$$\varlimsup_{n\to\infty} \frac{1}{n}\sum_{i\in A} n_i \leqslant \lim_{n\to\infty} \frac{1}{n}\Big(n - \sum_{i\in A_{k'}} n_i\Big) = 1 - \sum_{i\in A_{k'}} p_i \,;$$

and this being true for all k, we have

$$\varlimsup_{n\to\infty} \frac{1}{n}\sum_{i\in A} n_i \leqslant 1 - \sum_{i\in A'} p_i = \sum_{i\in A} p_i \,. \tag{14}$$

The assertion (9) follows from (13) and (14). Thus, in (9), the limit to the left exists and is given by the right-hand side of (9). We denote by $p(A)$ the probability of belonging to the new label A, namely $\lim_{n\to\infty} \Sigma_{i\in A}\, n_i/n = p(A)$, and write (9) also in the form

$$p(A) = \sum_{i\in A} p_i \tag{9'}$$

for any subset A of the discrete label space S.

By arguments similar to those above, used to show the insensitivity of $p(A)$ against a place selection in the case of a finite set A, it may be proved here that the new label A appears randomly in K if K_0 was a collective. That means that if a place selection is applied to K the limiting frequency in the selected subsequence K' exists and has the value $p(A)$.[3]

The mixing rule may now be summarized as follows:

If in a collective K_0 with discrete label space S the labels are "mixed" in such a way that the labels belonging to any subset A of S are replaced by a single label value, then in the new collective K the probability of this new label value is the sum of all probabilities p_i, $i \in A$, where the p_i are the original probabilities in K_0.

[3] In the proof of (9') randomness has not been used; if the p_i are merely chances then $p(A)$ is a chance and (9') holds.

Another way of summarizing our result with the emphasis on K_0 rather than on K is as follows:

Let $K_0 = K_0(G, S)$ *be a collective with probability distribution* $\{p_i\}$ *and denote by* $\mathfrak{S}$ *the set of all subsets of* S. *Then, for each* $A \in \mathfrak{S}$ *the frequency* $\sum_{i \in A} n_i/n$ *of elements of* A *converges with increasing* n *to the limit given by* (9'), *and this limiting value is insensitive to place selections of* G.

7.3. *Probability as a completely additive set function.* We see that if a collective $K_0(G, S)$ is given, it is possible to assign a probability $p(A)$ to *any* subset A of S, viz., to any $A \in \mathfrak{S}$. The probability of $A \in \mathfrak{S}$ is given by (9') and the so-defined probability distribution $p(A)$ is seen to be a non-negative additive set function defined on all sets of $\mathfrak{S}$ (see p.21). The collective K_0 may thus be denoted by $K_0(G, \mathfrak{S})$ rather than $K_0(G, S)$.

We wish to follow up this idea and therefore remind the reader of a few definitions, some of which have, in fact, already been used occasionally: a set $\mathfrak{S}$ whose elements are sets is called a *system* or a *family* of sets. The *union* or *sum* $A \cup B$ or $A + B$ of two sets A and B is the set of all elements which belong to at least one of the two sets; the *intersection* or *product* $A \cap B$ or AB is the set of all elements which belong to both sets. If A and B have no element in common, their intersection is the *zero set* or *void set* or *empty set*, which contains no element: $AB = 0$ in this case and A and B are called *disjoint* sets. The difference $A - B$ will be defined only if A contains each element of B and, in this case, it is the set of all elements which belong to A but not to B. A system π of sets is called a *field* if, together with any two sets A and B of π, the sum $A \cup B$, the intersection AB, and, the difference $A - B$ if it exists belong to π.[4] We call π a σ-*field* (sigma field) or *Borel field* provided that for every denumerable set $\{A_n\}$ of elements of π the union $A_1 + A_2 + \cdots$ (or $\bigcup_{n=1}^{\infty} A_n$) also belongs to π. It can be shown that a σ-field also contains the intersection $A_1 A_2 A_3 \cdots$ of any countable collection of sets in π.

We return to the last statement on K_0 and observe that *the set* $\mathfrak{S}$ *of all subsets of* S *is a Borel field.* In fact, the union (finite or denumerable) of elements of $\mathfrak{S}$ is an element of $\mathfrak{S}$ (we agree that $\mathfrak{S}$ contains the void set) and S itself is an element of $\mathfrak{S}$. On the Borel field $\mathfrak{S}$ we have defined a non-negative function $p(A)$ with the property $p(S) = 1$. It is

[4] Actually, it is not necessary to state explicitly that the intersection AB belongs to a field π. It is sufficient to postulate that together with two sets A and B, their sum and their difference belong to π, since $AB = B - [(A + B) - A]$.

clear from (9') that if A_1 and A_2 are disjoint subsets of S, then $p(A_1 \cup A_2) = p(A_1) + p(A_2)$. And if A_1, A_2, A_3, ... are pairwise disjoint subsets of S, then $p(A_1 \cup A_2 \cup A_3 \cdots) = p(A_1) + p(A_2) + p(A_3) + \cdots$. A function $p(A)$ which has this property is a *completely additive set function.*[5] *The probability defined by our assumptions* (1), (3) *is thus a completely additive set function over a* σ*-field* $\mathscr{S}$. [Note that with Eq. (3) assumed, the assertion (9) or (9') of complete additivity was *proved.*] The mixing rule on p. 00 is a statement on frequencies in infinite sequences of observations and not merely the expression of additivity of a set function.

We shall return to this point (see Appendix Two and Chapter II). At the moment we note only that additive set functions are encountered in a wide variety of applications. A mass distribution in mechanics is, for example, given by a non-negative additive set function. However, if we add to this fact some postulates regarding the differentiability of the coordinates with respect to time, etc., we by no means obtain the "axioms of mechanics" and we could not solve a single problem of mechanics. Similarly, it is necessary to formulate the specific features of probability distributions, those which distinguish them from mass distributions and other additive set functions (area, weight, electric charge, etc.) and to show their decisive role in the development of the theory.

Returning to the more formal considerations we turn our attention to an *n-dimensional discrete label space.* This is hardly a generalization. If we consider the simultaneous tossing of n distinguishable dice (see footnote 2, p. 3) the label space contains 6^n points and it contains k^n points in case of "k-valued dice." Also, in case of a countable label, the simultaneous observation of an n-tuple of results does not lead beyond the discrete label space. The collective consists then of an infinite sequence whose elements are n-tuples formed by the numbers 1, 2, ..., k or 1, 2, 3, ..., respectively. Accordingly we have, for example in the case of the dice, 6^n probabilities $\lim_{n\to\infty} N_x/N$, where x stands for x_1, x_2, ..., x_n, and each $x_i = 1, 2, ..., 6$ for $i = 1, 2, ..., n$. We note that the n dice may be "linked" in any arbitrary way. The randomness applies to the *succession of the labels* x *in the infinite series* $\{x_j\}$. It is obvious how this generalizes to the countable case: in this case the probability distribution is given by $p(x_1, x_2, ..., x_n)$, $x_i = 1, 2, 3, ...$ for $i = 1, 2, ..., n$.

The label space may now be denoted by S_n. All subsets of S_n form

[5] It is called an additive set function if this property holds for a finite number of sets A_i only.

a σ-field which we denote by $\mathcal{S}_n$. It is then clear that we have $\mathcal{S}_1 = \mathcal{S}$ and

$$\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_3 \subset \cdots \subset \mathcal{S}_n \subset \cdots .$$

Hence $\mathcal{S}_n$ may also be considered as the field of the subsets of all $\mathcal{S}_\nu$ for $\nu \leqslant n$. Finally, the field of *all* $\mathcal{S}_n$ may be denoted by T and we have

$$T = \bigcup_{\nu=1}^{\infty} \mathcal{S}_\nu$$

as the general field of the discrete collective. It is easily seen that while all $\mathcal{S}_n$ are σ-fields, T is a field but not a σ-field. T contains countably many sets.

Problem 7. A blue and a red die are simultaneously thrown and the probabilities of each of the 36 combinations of spots x and y are known to be equal. Find the probabilities of the possible products xy.

Problem 8. If in the preceding example, the probabilities have values p_{11}, p_{12}, p_{21}, ..., what is the probability of obtaining a result for which $x^2 + y^2 \leqslant 12$?

8. Third Operation: Partition

8.1. *Definition.* Consider a collective K_0 consisting of the sequence $x_1, x_2, x_3, \ldots$ of label values, where each x_i is one of the elements of the label set S. Let A be any subset of S, $A \subset S$, comprising finitely or infinitely many labels. We derive a new infinite sequence K by retaining only those elements of K_0 which belong to A and discarding all other elements. This operation, which we shall call a *partition*, is obviously not a place selection since the decision to retain or reject an element of K_0 depends on the (label) value of just this element.

Now let B be a subset of A, and denote by $n(A)$ and $n(B)$ the number of elements among the first n elements of K_0 which belong to A and B, respectively.[1] From the assumptions made under the mixing rule, it follows that the limits of the frequencies of A and B in K_0, namely,

$$\lim_{n\to\infty} \frac{n(A)}{n} = p(A), \qquad \lim_{n\to\infty} \frac{n(B)}{n} = p(B) \tag{15}$$

exist. The frequency of B in K is $n(B)/n(A)$. Now

$$\frac{n(B)}{n(A)} = \frac{n(B)}{n} \div \frac{n(A)}{n} .$$

[1] The change in notation, $n(A)$ rather than n_A, is necessary to avoid confusion; see Eq. (16) ff.

We denote[2] by $p_A(B)$ the probability—if it exists—of encountering a label B in the collective K. From the preceding equation and Eq. (15), if $p(A) \neq 0$, it follows that the limit

$$\lim_{n\to\infty} \frac{n(B)}{n(A)} = p_A(B) = p(B|A) \tag{16}$$

exists, and

$$p_A(B) = \frac{p(B)}{p(A)} \qquad \text{and} \qquad p_A(A) = 1. \tag{16'}$$

In terms of the collective K_0, $p_A(B)$ is the probability (or chance) of the label B if we know that some label out of A has been obtained. If $p(A) = 0$, then $p(B) = 0$ also. In this case $p_A(B)$ may still exist but cannot be expressed in terms of the probabilities (chances) $p(A)$, $p(B)$.

We observe that if the labels contained in A are a_1, a_2, a_3, ..., then by the mixing rule

$$p_A(a_1) + p_A(a_2) + p_A(a_3) + \cdots = \frac{p(a_1) + p(a_2) + p(a_3) + \cdots}{p(A)} = 1; \tag{17}$$

and in the collective K of label space A, the distribution consists of $p_A(a_1)$, $p_A(a_2)$, ... with sum equal to one. Each $p_A(a_i)$ is obtained from $p(a_i)$ by multiplication with the constant factor $1/p(A)$. Likewise if A is the sum of disjoint sets B_1, B_2, B_3, ..., then

$$p_A(B_1) + p_A(B_2) + \cdots = 1. \tag{17'}$$

We must still show that $p_A(B)$ is a probability rather than a chance, i.e., that it is unaffected by place selections. Assume, for example, that only those elements of K are retained which are immediately preceded by three elements with label value a_2; by this selection a sequence K' will result. We shall show that K' can also be obtained by partitioning a sequence K_0' which has been obtained from K_0 by a certain place selection, the definition of which is as follows: an element of K_0 is retained if, and only if, it is immediately preceded by a group of labels that includes the label a_2 three times, and does not include any other label belonging to A. If then in this sequence K_0' a partition with respect to A is carried out, the retained elements are identical with those elements of K which are preceded by an a_2-triplet. Thus we have again obtained K'.

If among the first n elements of K_0' a specific element a_i of A appears n_i' times while there are $n'(A)$ elements with labels belonging to the

[2] In exactly the same sense as $P_A(B)$, the notation $p(B \mid A)$ is also used.

subset A, the frequency of a_i after the partition—that is, the frequency of a_i in K'—is the limit of

$$\frac{n_i'}{n'(A)} = \frac{n_i'}{n} \div \frac{n'(A)}{n} .$$

The limits of n_i'/n and of $n'(A)/n$ are the same as the limits of n_i/n and $n(A)/n$ since K_0 is a collective; thus the limiting frequency of a_i in the sequence K' is again $p(a_i)/p(A)$. From this result for an arbitrary a_i in A we pass to the randomness of $p_A(B)$ by mixing those labels a_i which form B.

The following simple example illustrates the particular type of problems solved by the partition rule. We cast a die (label values 1, 2, ..., 6) and ask: what is the probability of the result "2" if it is known that an even number has been obtained? Here the subset A consists of the even numbers 2, 4, 6 and, according to the mixing rule, $p(A) = p_2 + p_4 + p_6$. Therefore, the probability in question is $p_A(2) = p_2/p(A) = p_2/[p_2 + p_4 + p_6]$. If all p_i are equal the result is $\frac{1}{6} \div \frac{1}{2} = \frac{1}{3}$.

Before proceeding we summarize as follows:

From a collective K_0 with discrete label space S including more than two points, one can derive another collective K by partitioning with respect to a subset A of S, retaining only those elements of K_0 which belong to A. The probability $p_A(B)$ of a subset B of A within the derived collective K equals the quotient $p(B)/p(A)$, where $p(A)$ and $p(B)$ are the probabilities of A and B, respectively, in K_0 and $p(A) \neq 0$. $p_A(B) = p(B|A)$ is also called the (conditional) probability of the result B if it is given that A has happened.

8.2. *Extensions and comments.* If the original sequence K_0 does not fulfill the condition of randomness and the a_i have chances only, one can still derive a sequence K by partitioning with respect to A. The frequency values in K exist and their limits have the values $p(B)/p(A)$—for subsets B of A—but they will then be chances rather than probabilities.

Thomas Bayes (1763) was probably the first to consider quotients of probabilities with a view toward conditional probabilities. Formula (16′) is sometimes referred to as Bayes' rule. We shall reserve this name for the application to certain problems that Bayes had in mind (Chapter VII).

The rule of partitioning is often applied to the following situation. Consider a subset A of S and derive K from K_0 as before; now, however, instead of a subset B of A we consider a second subset C of S, which need not be a subset of A. The intersection $AC = B$ will be either void or a subset of A for which (16′) holds. The sample space of K is A,

which does not contain any labels of $C - B$; labels belonging to $C - B$ appear in K with frequency zero (for every value of n); hence, $p_A(C - B) = 0$ is the chance, in K, of a label belonging to $C - B$. Then we define

$$p_A(C) = p_A(B) + p_A(C - B) = p_A(B) + 0 = h_A(AC) \tag{18}$$

as the probability (chance) of finding in K a label of C. To $p_A(AC)$ formula (16′) applies and we obtain

$$p_A(C) = p_A(AC) = \frac{(pAC)}{p(A)}. \tag{19}$$

$p_A(C)$ is again called the conditional probability of the set (event) C if it is given that A has happened. We illustrate by an example. The label space S consists of the 6 labels 1, 2, 3, 4, 5, 6; let $A = 1, 2$ and $C = 2, 4, 6$; AC is the label 2; then $p_A(C)$, the probability of an even result if we know that the result was either 1 or 2, equals $p_A(C) = p(2)/[p(1) + p(2)]$, where we now write $p(i)$ instead of p_i.

Using (19) we obtain the following formula. Let $A_1, A_2, \ldots$ be a number of disjoint sets such that $\Sigma A_i = S$ and $p(A_i) > 0$. If A is a subset of S, then

$$p(A) = p(A_1)p_{A_1}(A) + p(A_2)p_{A_2}(A) + \cdots. \tag{19′}$$

The simple proof is left to the reader.

We consider a second example. Let K_0 refer to the repeated simultaneous tossing of two distinguishable dice I and II. The label space S consists of the 36 points (x, y), $\begin{matrix} x \\ y \end{matrix} = 1, 2, \ldots, 6$, and the corresponding distribution is $p(x, y)$ with $\Sigma_x \Sigma_y p(x, y) = 1$. Let A be the subset consisting of the six points $(3, y)$, $y = 1, 2, \ldots, 6$, that is, the event: "first die shows 3"; and C the set $(x, 5)$, $x = 1, 2, \ldots, 6$, that is, the event: "second die shows 5." Here C is not a subset of A (but is, of course, a subset of S) and we apply (19). The intersection of A and C is the point $(3, 5)$ and $p_A(C) = p''(5|3)$ is the conditional probability that the second die shows 5 if the first die shows 3. The double prime in $p''(5|3)$ is to remind us that we consider a probability for the second die. We have

$$p''(5|3) = \frac{p(3, 5)}{p(3, 1) + \cdots + p(3, 6)} = \frac{p(3, 5)}{p'(3)},$$

where $p'(3)$ is the probability of the result 3 for the first die. Likewise we denote by $p'(x) = p(x, 1) + \cdots + p(x, 6)$, $x = 1, 2, \ldots, 6$ [and by

$p''(y) = p(1, y) + \cdots + p(6, y)$, $y = 1, 2, \ldots, 6$] the *marginal probabilities* to obtain the result x (the result y) with the first (the second) die. Hence, we have the formulas

$$p''(y|x) = \frac{p(x, y)}{p'(x)} \qquad \text{and similarly} \qquad p'(x|y) = \frac{p(x, y)}{p''(y)}, \tag{20}$$

with

$$\sum_y p''(y|x) = 1, \qquad \text{and} \qquad \sum_x p'(x|y) = 1. \tag{20'}$$

We stress that formulas (20) presuppose the knowledge of the two-dimensional collective K_0 with distribution $p(x, y)$, $\begin{matrix} x \\ y \end{matrix} = 1, 2, \ldots, 6$. Note also that formulas (20) also follow from (16) with A the same set as before, consisting of the six points $(x, 1), \ldots, (x, 6)$, and B the one point (x, y), a subset of A.

Problem 9. At a streetcar stop, the trains of five lines pass with the probabilities $p_1, p_2, \ldots, p_5$ (sum $= 1$). For the first three lines two-car trains are in use. If a man waiting for a line-2 train sees a train with two cars approaching, what are the chances that it is his train?

Problem 10. It is known that a throw of a certain die with probabilities $p_1, p_2, \ldots, p_6$ resulted in a prime number. What is the probability that this number was smaller than four?

Problem 11. A bag contains lots numbered 1 to 90. The probability of drawing any odd number is p' and of drawing any even number is p'', where $p' = 1.2\,p''$. If it is known that a multiple of 5 was drawn, what is the probability that that number is 25? How would the result change if the relation between p' and p'' were $p'' = 1.2\,p'$?

9. Fourth Operation: Combining

9.1. *Combination of collectives derived by related place selections.* In the three basic operations discussed hitherto, one single collective K_0 served each time as point of departure for the construction of a new collective. Now we shall consider the problem of combining two or more given collectives. We start with a very special example.

In the tossing of a single die the customary label set consists of the points 1, 2, ..., 6. Each of these has a certain probability $p_1, p_2, \ldots, p_6$, with the sum equal to one. From K_0 we derive two other collectives, K_1 and K_2, by two different place selections, retaining in K_1 all elements of the original sequence with odd ordinal number and in K_2 those with even ordinal number. According to the rule of place selection,

the probabilities of the six labels will still be p_1, p_2, ..., p_6 in K_1 as well as K_2. Now we derive from K_1 and K_2 a new sequence K by *combining* the elements of both in the following way. The first element of K consists of the first elements of K_1 and K_2, that is, of the results of the first and second tossings in the original sequence; likewise the ith element is the combination of the results of the two tossings numbered $2i - 1$ and $2i$, that is, a pair of digits out of the set 1, 2, ..., 6 (for instance, the pair 3, 5). Thus the label set of K consists of the 36 points whose coordinates are the integers 1, 2, ..., 6. We ask whether the sequence K is a collective, and, if so, what the probabilities of the 36 label values are.

Let us investigate the frequency of the point (3, 5) in K. Among the first n elements of K_1 there will be n_3 elements showing the result 3, the limit of n_3/n being p_3. Now, we consider the following place selection applied to the original collective K_0: retain those elements whose ordinal number is even and which are immediately preceded by a term of label value 3. Now take the first n_3 elements of this selected subsequence and count the number of labels 5 among them. Let this number be $n_{3,5}$; the limit of $n_{3,5}/n_3$ with n_3 increasing to infinity, is p_5 since we have merely performed a place selection. With respect to K, the significance of the numbers n_3 and $n_{3,5}$ is this: among the first n elements of K there are n_3 whose first component is 3 and $n_{3,5}$ whose first component is 3 *and* whose second component is 5. Therefore, the frequency of the combination (3, 5) is

$$\frac{n_{3,5}}{n} = \frac{n_{3,5}}{n_3} \cdot \frac{n_3}{n}. \tag{21}$$

As n goes to infinity, the first ratio to the right approaches p_5 and the second p_3. Thus we see that the chance of the result (3, 5) exists in K and that it equals the product $p_3 p_5$, or more generally, using $p(a)$, $p(a, a')$:

$$p(a, a') = p(a)\, p(a') \qquad \left(\begin{matrix} a \\ a' \end{matrix} = 1, 2, \cdots, 6\right). \tag{22}$$

It can easily be seen that the sequence K also fulfills the condition of randomness. In fact, a place selection in K entails a place selection in K_0, transforming this sequence into K_0'. By using K_0' in the way K_0 was used before, one sees that the effect of a place selection in K is identical with that of a certain place selection in K_0. Thus $p(a, a') = p(a)\, p(a')$ is a genuine probability.

We showed that in K_0 the probability of any given pair of labels (a, a') exists and equals $p(a)\, p(a')$ where the ordinal number of the first label,

a, was odd, that of a' even. Now the same is true for the pairs such that the ordinal number of the first label, a, is even, that of a' odd. If we consider all (overlapping) pairs of successive elements, then in half of them the first element will have an odd, in the other half an even ordinal number. If follows that the probability of the succession (a, a') in K_0 is $\frac{1}{2}p(a)p(a') + \frac{1}{2}p(a)p(a') = p(a)p(a')$. We have the result:

(1) *In a collective K_0 the probability of every pair of successive labels exists and its value is the product of the original probabilities which constitute the pair.*

It is easy, although lengthy, to extend the preceding operations to the case where the collective K_0 is defined with respect to a denumerable label set S with probability distribution p_i, $i = 1, 2, \ldots$ and where more general place selections are used to define K_1 and K_2.

Suppose that $\{\alpha_i\}$, $\{\beta_i\}$ are both sequences of increasing integers and that the inequalities

$$1 \leqslant \alpha_1 < \beta_1 < \alpha_2 < \beta_2 < \cdots < \alpha_i < \beta_i < \cdots \tag{23}$$

are satisfied. We shall call the two place selections defined by $\{\alpha_i\}$, $\{\beta_i\}$ *related* to each other. (In the former example we had $\alpha_1 = 1$, $\alpha_2 = 3$, $\alpha_3 = 5, \ldots$, $\beta_1 = 2$, $\beta_2 = 4$, $\beta_3 = 6, \ldots$) Let K_1 and K_2 be the subsequences of K_0 selected by $\{\alpha_i\}$ and $\{\beta_i\}$, respectively, and form a new sequence K by combining in pairs the elements numbered α_i, β_i of K_1 and K_2:

$$K: \quad (x_{\alpha_1}, x_{\beta_1}), (x_{\alpha_2}, x_{\beta_2}), \ldots, (x_{\alpha_i}, x_{\beta_i}), \ldots . \tag{24}$$

Consider any two elements a and a' (not necessarily distinct) of the label space S. *We shall show that the element (a, a') appears in K with probability $p(a)p(a')$.*

Suppose that among the first n pairs of K there are n_a whose first entry is a and $n_{a,a'}$ whose second entry is a' and whose first entry is a. Now, n_a is clearly also the number of times a appears among the first n elements of K_1 so that $\lim_{n\to\infty} n_a/n = p(a)$. Next we show how this same number, $n_{a,a'}$, is obtained by means of a place selection applied to K_0: we single out in K_0 each element whose ordinal number is a β_i provided the corresponding x_{α_i} has the value a and call this sequence K_0'; now take the first n_a elements of K_0' and count the number of those among them which have the value a'; this number is exactly $n_{a,a'}$. Since K_0' is a collective, it follows that $\lim_{n_a\to\infty} n_{a,a'}/n_a = p(a')$. Referring again to K, we see that

$$\begin{aligned} \lim_{n\to\infty} n_{a,a'}/n &= \lim_{n\to\infty} (n_{a,a'}/n_a \cdot n_a/n) \\ &= \lim_{n\to\infty} n_a/n \lim_{n\to\infty} n_{a,a'}/n_a \\ &= p(a)p(a'). \end{aligned} \tag{25}$$

This shows that the chance of a given pair (a, a') exists in K and that its value is the product of the chances (in K) of a and a'.

In order to show that K is a random sequence, we subject K to a place selection Γ which produces the sequence of pairs

$$K_\gamma: \quad (x_{\alpha_{\gamma_1}}, x_{\beta_{\gamma_1}}), (x_{\alpha_{\gamma_2}}, x_{\beta_{\gamma_2}}), \ldots, (x_{\alpha_{\gamma_i}}, x_{\beta_{\gamma_i}}), \ldots . \tag{26}$$

We have to show that in K_γ the limiting frequency of any given pair (a, a') exists and has the value $p(a)\,p(a')$. To see this we turn again to K_0 and subject it to the place selection

$$\alpha_{\gamma_1} \quad \beta_{\gamma_1}, \quad \alpha_{\gamma_2}, \quad \beta_{\gamma_2}, \ldots, \alpha_{\gamma_i}, \quad \beta_{\gamma_i}, \ldots,$$

and obtain a collective K_0' with the same distribution as K_0. From K_0' we derive two collectives K_1' and K_2' by the related place selections

$$\begin{aligned} K_1': &\quad x_{\alpha_{\gamma_1}}, \quad x_{\alpha_{\gamma_2}}, \quad x_{\alpha_{\gamma_3}}, \ldots \\ K_2': &\quad x_{\beta_{\gamma_1}}, \quad x_{\beta_{\gamma_2}}, \quad x_{\beta_{\gamma_3}}, \ldots, \end{aligned} \tag{27}$$

and we combine K_1' and K_2' to form the sequence of pairs $(x_{\alpha_{\gamma_i}}, x_{\beta_{\gamma_i}})$, $i = 1, 2, 3, \ldots$. This, however, is precisely the sequence K_γ which we had previously defined by means of a place selection on K. Thus K_γ has been derived by the above steps—K_0, K_0', K_1', K_2'—exactly as K was formed from K_0, K_1, K_2. We see by the same reasoning as in the proof of (24) that the chance of (a, a') in K_γ is $p(a)p(a')$. Therefore, K is a collective.

(1′) *If two collectives K_1 and K_2 are derived from a collective K_0 by related place selections* (23), *the combination K of K_1 and K_2 is again a collective and the probability of any label combination x, y in K is*

$$p(x, y) = p(x)\,p(y), \tag{25′}$$

where $p(x)$, $p(y)$ are the probabilities of x and y, respectively, in K_0.

It is easily seen that properties analogous to (1) and (1′) can be established for three labels, or for four, etc.

The here-proved multiplicative property of a random sequence is certainly its most obviously important property. The question poses itself whether it might be appropriate to characterize a random sequence by postulating directly some "multiplicative law within the sequence" instead of the randomness defined as insensitivity to place selections. It seems, however, that general randomness cannot be obtained in this way except by introducing some very artificial multiplicative property.

The statements (1) and (1′) form our first statements on the *multiplication of probabilities*. They are based entirely on the randomness of K_0.[1] Two more propositions will be derived which are partly independent of the randomness assumption.

[1] In the regular sequence K_0: 0, 1, 0, 1, 0, 1, ... with $p_2 = p_1 = \frac{1}{2}$, the four possible pairs (0, 0), (0, 1), (1, 0), (1, 1) have clearly the chances $p(0, 0) = p(1, 1) = 0$ $p(1, 0) = p(0, 1) = \frac{1}{2}$. If we perform the related place selections defined by $\{\alpha_i\} = 1, 5, 9, 13, \ldots, \{\beta_i\} = 3, 7, 11, 15, \ldots$ then $p(0, 0) = 1$, all others zero. In a collective with $p_1 = p_2 = \frac{1}{2}$, all $p(i, k) = \frac{1}{4}$.

9.2. *The multiplication rule. Independent collectives.* Assume that two dice are tossed simultaneously, the two sequences being K_1 and K_2. Let the probabilities in K_1 be $p_1', p_2', \ldots, p_6'$ and those in K_2 be $p_1'', p_2'', \ldots, p_6''$. By the combination of corresponding tossings we obtain a sequence K of elements with a two-dimensional label x, y and a label set of 36 points. Among the first n elements of K there will be n_x' elements with the first label component equal to x, and $\lim_{n\to\infty} n_x'/n = p'(x)$. Among these n_x' elements, there will be a number, say $n_{x,y}$ whose second component is equal to y. Thus, the frequency of the combination x, y is

$$\frac{n_{x,y}}{n} = \frac{n_x'}{n}\frac{n_{x,y}}{n_x'}. \tag{28}$$

The first factor to the right has the limit p_x', but what about the second one? The $n_{x,y}$ results have been obtained out of the complete sequence K_2 by a new type of selection: we chose a label x of K_1, for example, $x = 3$, and selected those results $y = 5$ of K_2 which occurred simultaneously with a result x in K_1. This is neither a place selection nor a partition. We shall say that K_2 *was sampled by means of the label* x *of* K_1. It may happen that this sampling does not affect the frequency limit in K_2, that is, *the limits of* $n_{x,y}/n_x'$, *whatever the* x, *are equal to one another and to the limiting frequency of the label* y *in* K_2,

$$\lim_{n\to\infty} n_{x,y}/n_x' = p''(y). \tag{29}$$

If this is the case for all x *and* y *we shall say that the collective* K_2 *is independent of* K_1. The limit of the second factor in (28) is then equal to the probability $p''(y)$ of the label y within K_2 and the limit of the right-hand side of (28) equals $p'(x)p''(y)$. We state:

(2) *If the collective* K_2 *is independent of* K_1 (*that is, if a sampling on* K_2 *performed by means of* K_1 *does not change the limiting frequencies in* K_2), *the combination of* K_1 *and* K_2 *leads to a sequence of elements in which the arbitrary label combination* x, y *has the limiting frequency*

$$p(x, y) = p'(x)\,p''(y). \tag{30}$$

We have introduced K_2 as independent of K_1. However, as seen from (30) in which $p'(x)$ and $p''(y)$ play symmetric roles, independence is a reciprocal property. In other words, *if* K_2 *is independent of* K_1, *then* K_1 *is also independent of* K_2; *they are mutually independent.* We consider the equation

$$\frac{n_{x,y}}{n} = \frac{n_y''}{n}\frac{n_{x,y}}{n_y''}, \tag{28'}$$

which is analogous to (28); and since we know from (30) that the $\lim_{n\to\infty} n_{x,y}/n$ exists and equals $p(x, y)$ and since $\lim_{n\to\infty} n_y''/n = p''(y)$, it follows that $\lim_{n\to\infty} n_{x,y}/n_y''$ exists and equals p_x': the sampling on K_1 performed by means of K_2 does not change the limiting frequencies in K_1. Thus the assumption of independence may be worded as follows: the frequency limits in the sampled sequences are independent of the label by means of which they are sampled.

Our considerations are not changed in any essential way if we admit any discrete label spaces S_1 and S_2 for K_1 and K_2, instead of the illustrative dice example, and replace the single labels x and y by arbitrary subsets A_1 and A_2 of $\mathfrak{S}_1$ and $\mathfrak{S}_2$ ($\mathfrak{S}_i$ being the set of all subsets of S_i including S_i itself, $i = 1, 2$). We obtain

$$p(A_1, A_2) = p'(A_1)\, p''(A_2), \qquad A_1 \in \mathfrak{S}_1, \quad A_2 \in \mathfrak{S}_2. \tag{30'}$$

We return to Eq. (30) and consider the question of randomness. It is not necessarily true that the frequency limit $p(x, y)$ in (30) is unaffected by a place selection perfomed on the combination sequence K. If this is so, $p(x, y)$ is a probability and we call K_1 and K_2 *combinable and independent*, or, more briefly, independent.

If we analyze[2] the meaning of a place selection performed on K, we find that in order to assure insensitivity of K to place selections we have to add the following restriction to the above definition of independence: K_1 and K_2 are combinable and independent if the values of the limiting frequencies in K_2 remain unchanged when we first make an arbitrary selection in K_1, then use any label in this partial sequence of K_1 for sampling K_2, and finally apply an arbitrary place selection to the sampled subsequence of K_2.

It is customarily assumed that in the simultaneous tossing of several dice under normal circumstances (use of one or more dice cups by one or more persons, etc.) any two sequences of single tossings are combinable and independent, thus giving rise to combined collectives in one or more dimensions. This assumption agrees with general experience as far as the results of a mathematical theory can be checked experimentally. In such a case we say briefly that *the multiplication rule* (30) *holds*.

If in the process of sampling (p. 30) K_1 is replaced by a mathematically defined infinite sequence of elements with existing limiting frequencies $p'(x)$, then the sampling performed on K_2 by means of K_1 becomes a place selection (of the type where the decision whether to retain or to discard an element depends on its ordinal number only). The insen-

[2] See R. v. Mises [22], p. 53.

sitivity of the frequency limits in K_2 is then simply a consequence of the assumed randomness of K_2 and (30) holds true.

9.3. *Dependent collectives.* Finally we consider the case where the quotient $n_{x,y}/n_x'$, which occurs in Eq. (28), is not independent of x in the limit. In terms of collectives this means that the collective K_2 is replaced by a set of collectives, each corresponding to one particular value of the variable x in K_1. In the dice example: there is a certain probability of casting 5 with the second die if the first die gives 3 and another probability of 5 with the second die if the first die shows 4, etc. Here is a more natural example of a probability problem of this type: An urn contains balls each marked with a number x, where x belongs to the set S: $(a_1, a_2, \ldots, a_k)$. The probability $p'(x)$ of drawing a ball with the number x (collective K_1) will depend on the ratio of balls numbered x in the urn to the total number of balls. [Under usual circumstances, it is assumed that $p'(x)$ is equal to this ratio.] Suppose now that after drawing one ball and *before returning it*, a second ball is drawn from the urn. The content of the urn for this second drawing now depends on the label x of the ball drawn at the first step. The chance of drawing a ball marked y in the second step will therefore be a function of y and x. It may be written as $p''(y \mid x)$, where the bar separating y from x serves to indicate that the probability of the single label value y depends on the parameter x. Both $p'(x)$ and $p''(y \mid x)$ are supposed to be known. The combination sequence K has a two-dimensional label, each element of K consisting of a pair of consecutive drawings with both balls returned after the second drawing only. Among the first n elements of K there will be n_x' with the first label x. Among these n_x' elements we shall have $n''_{y|x}$ elements in which the second label is y. The frequency of elements labeled (x, y) will then be

$$\frac{n''_{y|x}}{n} = \frac{n_x'}{n} \cdot \frac{n''_{y|x}}{n_x'} \tag{31}$$

and the limits of the two quotients to the right are $p'(x)$ and $p''(y \mid x)$, where

$$\sum_x p'(x) = 1, \qquad \sum_y p''(y \mid x) = 1. \tag{32}$$

If the limits $p'(x)$ and $p''(y \mid x)$ are insensitive to a place selection operated on K we say that the collective K_1 and the set of collectives $K_2(x)$ are *combinable* or: combinable but interdependent. More explicitly: we have to replace K_2 by $K_2(x)$ in the previous explanation (p. 31) of

"combinable." This is certainly the case if K_1 is replaced by any mathematically defined sequence with limiting frequencies $p'(x)$.

The reader will have no difficulty in generalizing the considerations of this paragraph to the case of denumerable label sets. We formulate:

(3) *If a collective K_1 with probabilities $p'(x)$ is combined with a set of collectives $K_2(x)$ for which the probabilities of a label y, namely, $p''(y \mid x)$ depend on the parameter x, the resulting sequence K has limiting frequencies*

$$p(x, y) = p'(x)\, p''(y \mid x). \tag{33}$$

If K_1 and $K_2(x)$ are combinable, $p(x, y)$ is a probability distribution.

By summation of Eq. (33) over all y we see that, on account of the second equation (32), $p'(x) = p_1(x)$, where $p_1(x) = \Sigma_y\, p(x, y)$ is the *first marginal distribution* of $p(x, y)$. Note that $p''(y \mid x)$ can be summed with respect to y but not with respect to x. A result similar to (33) and with an analogous meaning is

$$p(x, y) = p''(y)\, p'(x \mid y), \tag{33'}$$

where

$$\sum_y p''(y) = 1, \qquad \sum_x p'(x \mid y) = 1 \tag{32'}$$

and $p''(y) = p_2(y) = \Sigma_x\, p(x, y)$ is the second marginal distribution.

We finish with a remark on the relation between combination and partition. Remember the partition formula (19) (p. 25) and the first formula (20). In terms of the dice problem of p. 26, $p''(y \mid x)$ was the conditional probability of the result y for the second die if we knew that x was the result for the first die:

$$p''(y \mid x) = \frac{p(x, y)}{p'(x)} = \frac{p(x, y)}{\Sigma_y\, p(x, y)}. \tag{20}$$

The starting point was the two-dimensional distribution $p(x, y)$, $x = 1, \ldots, k$, $y = 1, \ldots, j$; the given data are the $kj - 1$ values $p(x, y)$, and we derive $p''(y \mid x)$ from them by partition.

In the problem of combining which led to (33) the starting point was not a two-dimensional distribution but one distribution $p'(x)$ for the first die and a set of k one-dimensional distributions $p''(y \mid x)$. [If $x = 1, \ldots, k$, $y = 1, \ldots, j$ the data are: k values $p'(1), \ldots, p'(k)$ with sum one and kj values $p''(y \mid x)$ where $\Sigma_y\, p''(y \mid x) = 1$ for each x, hence altogether $(k - 1) + (kj - k) = kj - 1$ values.] From these data we derive by (33) the two-dimensional distribution $p(x, y)$.

If we partition after combining we wind up again where we started (and conversely). An actual situation might be analyzed in any of several ways.

Problem 12. In a game of chance the three participants have the probabilities p_1, p_2, p_3 to win, with $p_1 + p_2 + p_3 = 1$. What is the probability that in a set of 5 independent games the first player wins (a) the first as well as the last of the five games and no other games, and (b) the first and the last games?

Problem 13. Assume that in the preceding example a set consists of three successive games (3-dimensional collective). What is the probability for the first player to win at least 2 out of the 3 games?

Problem 14. On casting a die with the probabilities $p_1, p_2, \ldots, p_6$, what is the probability that the difference between two consecutive results is ± 3?

Problem 15. Bertrand's box problem. We have boxes of identical appearance each of which has two drawers and in each drawer either a gold or a silver coin. There are three types of boxes:

(1) containing 2 gold coins,

(2) containing 2 silver coins,

(3) containing 1 silver and 1 gold coin.

In selecting a box, the chances are p_1, p_2, p_3 for getting one of type 1, 2, 3, respectively. With a box of type 3 the chance of opening the drawer with the gold coin is q_1, of opening the drawer with the silver coin $q_2 = 1 - q_1$. If a box has been selected, a drawer opened and a gold coin found, what is the probability that the other drawer of this box contains

(a) a gold coin, (b) a silver coin?

Problem 16. Study the sequence (b) of p. 8. Denote by $v(x)$ the chance of the argument x; by $v(x \mid y)$ the chance that x comes after y; by $v(x, y)$ the joint chance of x and y in succession, where $x, y = 0, 1, 2$. Prove that:

(1) $v(0) = v(2) = \frac{1}{4}, \quad v(1) = \frac{1}{2}.$

(2) $v(0 \mid 0) = \frac{1}{2}, \quad v(1 \mid 0) = \frac{1}{2}, \quad v(2 \mid 0) = 0$
$v(0 \mid 1) = \frac{1}{4}, \quad v(1 \mid 1) = \frac{1}{2}, \quad v(2 \mid 1) = \frac{1}{4}$
$v(0 \mid 2) = 0, \quad v(1 \mid 2) = \frac{1}{2}, \quad v(2 \mid 2) = \frac{1}{2}.$

(3) $v(0, 0) = v(0, 1) = v(1, 0) = v(1, 2) = v(2, 1) = v(2, 2) = \frac{1}{8},$
$v(1, 1) = \frac{1}{4}, \quad v(0, 2) = v(2, 0) = 0.$

Compare your results with the corresponding ones in the sequence of results of tossing independently two coins (1 = heads, 0 = tails) with the sum of the two coin values as label.

10. Additional Remarks on Independence

10.1. *Review.* In this section we wish to comment on a rather common definition of independence which seems unsatisfactory to us. In order to have a firm basis for the discussion, we start by reviewing (partly) our concept of independence, as obtained either from the partition formula (20) applied to a given two-dimensional $p(x, y)$ or from the combination formula (33) where $p(x, y)$ was obtained as the result of combining.

Starting with (33) we arrive at "independence" if we assume that the k distributions $p''(y \mid x)$ are identical, i.e., that they are equal to the same distribution $p''(y)$ for all values of the parameter x. In terms of the dice problem, that means that while we are tossing the red die, there is one and only one distribution for the black die, no matter what result appears for the red one. Then (33) becomes

$$p(x, y) = p'(x)\, p''(y), \tag{36}$$

and the remarks about randomness on p. 31 apply to the collective with distribution $p(x, y)$. Note that $p(x, y)$ as obtained in (27) has in the discrete case a matrix of rank one, since clearly $p(x, y)p(u, v) = p(x, v)p(u, y)$ for all combinations of the arguments.

Next, we take the partition (20) as our starting point. We ask whether, under certain circumstances, the conditional probability $p''(y \mid x) = p(x, y)/\Sigma_y\, p(x, y)$ is the same for all x, thus depending on y only. It makes sense to ask whether, in certain cases, the influence of x cancels out, so that, in terms of the dice example

$$p''(y \mid 1) = p''(y \mid 2) = \cdots = p''(y \mid 6), \qquad \text{for all } y, \tag{34}$$

or, equivalently, with $p'(x) = \Sigma_y\, p(x, y)$, whether, for all y:

$$\frac{p(1, y)}{p'(1)} = \frac{p(2, y)}{p'(2)} = \cdots = \frac{p(6, y)}{p'(6)}\,. \tag{34'}$$

If, in (34') we replace $p'(1)$, $p'(2)$, ... by their definition we see that Eqs. (34') hold if

$$\begin{aligned} &p(1, y)p(2, 1) - p(2, y)p(1, 1) = 0 \\ &p(1, y)p(2, 2) - p(2, y)p(1, 2) = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots, \text{ etc.,} \end{aligned} \tag{35}$$

for all y, that is, if the rank of $p(x, y)$ equals one. If (34′) and therefore (34) hold, $p''(y \mid x)$ is then a function of y alone which we denote by $p''(y)$:

$$p(x, y) \Big/ \sum_x p(x, y) = p''(y). \tag{34''}$$

Then, from (34″)

$$p(x, y) = p'(x)p''(y), \tag{36}$$

and $p(x, y)$ is seen to be factorable. As before, from (36)

$$p'(x) = \sum_y p(x, y), \qquad \text{and} \qquad p''(y) = \sum_x p(x, y) \tag{37}$$

follows. If $p(x, y)$ in (36) is, by hypothesis, a probability (in contrast to a chance), then $p'(x)$, $p''(y)$ are probabilities since they have been obtained from $p(x, y)$ by mixing. Using the notion of rank, we may thus say: If the matrix of a discrete two-dimensional probability distribution $p(x, y)$ is of rank one, then and only then $p(x, y)$ factors into the product $p'(x)p''(y)$ of two independent collectives.

Note that whether we take partitioning or combining as our starting point, the mathematical and physical conditions for independence, in our sense, concern *the interconnection* of the two dice, of the two one-dimensional collectives, of the two "random variables" x and y, or, in terms of $p(x, y)$, the *type of this two-dimensional distribution* (rank one, factorable); they are not concerned with properties of each single die.

10.2. *An unsatisfactory definition.* We return to Eqs. (19) and wish to discuss the way in which it is often used as a starting point for the definition of "independence." Let A and C be two arbitrary subsets of a label space S, with AC denoting the intersection of A and C. One defines independence by the condition that $p_A(C) = p(AC)/p(A)$ be equal to $p(C)$ or

$$p(A)p(C) = p(AC). \tag{38}$$

Since we disagree with this undiscerning use of Eq. (19) and since independence is a central and elementary concept of probability theory, we wish to consider the problem carefully. Consider the following example: The label space S consists of the six points 1, 2, 3, 4, 5, 6 with distribution p_i, $i = 1, \ldots, 6$; the event (or set) A consists of the three points 2, 3, 4, the event C of the two points 1, 2; the intersection AC is the point 2, and $p_A(C) = p_2/(p_2 + p_3 + p_4)$. All this is, of course,

correct. Now, however, the following question is asked: Under what conditions is $p_A(C)$ equal to $p(C)$? In our example

$$p_2/(p_2 + p_3 + p_4) \stackrel{?}{=} p_1 + p_2.$$

The example is so chosen that this is true for $p_i = \frac{1}{6}$, $i = 1, \ldots, 6$. The statement is then made that, in this case, the events A and C are independent. Let us analyze this statement.

Let us consider a set A consisting of the points 2, 3, 4 and a set C of point 2; here $C \subset A$. Then $p_A(2) = p_2/p_A = p_2/(p_2 + p_3 + p_4)$. Here $p_A(2) \equiv p(2 \mid A)$ certainly does not remain unchanged if we vary the set A, the way we varied the x in $p''(y \mid x)$ in Eqs. (34), and certainly for no A, $p(A) \neq 1$, is $p_A(2)$ equal to p_2. Now, however, in order to make such an equality possible, one considers other sets, C, such that the intersection AC is again the label 2 but $C \supset AC$. Such subsets of our S are, for example, $C_1 = (1, 2)$, $C_2 = (2, 5)$, $C_3 = (1, 2, 5, 6)$, $C_4 = (1, 2, 5)$. Then, for each of these C_i

$$p_A(C_i) = p(AC_i)/p(A) = p_2/(p_2 + p_3 + p_4).$$

Thus, having the choice of sets C_i one may ask whether for one or more of them, and with some given distribution, $p_A(C) = p(C)$. If all $p_i = \frac{1}{6}$, this holds true for $C_1 = (1, 2)$ or for $C_2 = (2,5)$ but not for C_3 or C_4. If we take $p_1 = p_5 = \frac{1}{12}$, $p_2 = p_3 = p_4 = \frac{1}{6}$, $p_6 = \frac{1}{3}$, then the above equality holds for $C_4 = (1, 2, 5)$ but no longer for C_1 or C_2, and so on. It seems that definition (38) allows the possibility of purely numerical accidents. What *is* the meaning of the statement that, for a given distribution, "the events $A = (2, 3, 4)$ and $C = (2, 5)$ are independent" while "the events (2, 3, 4) and (1, 2, 5) are dependent" or "events (1, 6) and (2, 3, 4) are dependent"?[1]

We have seen, however, that one can arrive at a meaningful concept of independence by way of partition. What is the difference between that procedure and the present one? A first immediate remark would be that independence should be defined for collectives rather than for isolated events. If, however, with any given $A \subset S$, $C \subset S$, we form the two

[1] The simplest example of this type, quoted by Kolmogorov, is a label set consisting of four points, 1, 2, 3, 4, each with $p_i = \frac{1}{4}$. Then, if $A = (1, 2)$, $C = (2, 3)$, $p(A) = p(C) = \frac{1}{2}$, $p(AC) = \frac{1}{4}$, the events (1, 2) and (2, 3) are called "independent."

This author's name appears as Kolmogoroff in his German publications. We have used the spelling Kolmogorov, the official American transliteration from the Russian, for consistency throughout this volume.

collectives, K_1 with labels A and $A' = S - A$, and K_2 with labels C and $C' = S - C$, and if, then, for a particular distribution

$$p(A)p(C) = p(AC), \tag{38}$$

then the three analogous conditions, for example $p(A')p(C) = p(A'C)$, hold automatically. We dismiss this objection.

A more serious point seems, however, to be the one briefly mentioned on p. 4. One may *say* that the intersection of two sets A and C has the "property" of belonging to A and the "property" of belonging to C (and to many others). Nevertheless, the label "2"—*the result of the ordinary tossing of one die*—is not a two-dimensional label[2] like "blond hair, blue eyes" or "first die 3, second die 5." In the case of Eq. (20) or (36) the label space of $p(x, y)$ is two-dimensional; hence it is natural to ask under what conditions the two-dimensional $p(x, y)$ factors into the product of two one-dimensional distributions; or to ask, under what conditions $p''(y \mid x)$ is independent of the parameter x and reduces to $p''(y)$.

In a meaningful concept of independence *two* "properties" are involved which may or may not influence each other, In contrast to that, a definition based on Eq. (38) (to which we did our best to give a meaningful semblance) remains a watered-down generalization of a meaningful concept.[3] We think there are two legitimate ways of arriving at independence, one, starting with the combination of two one-dimensional collectives, the other, with the partitioning of a two-dimensional collective. But the fact that a set is regarded as the intersection of two (or of twenty) other sets does not make it the label of a two-dimensional collective. Always, $p(AC)$ is the measure of a set; but unless it is a two-dimensional probability with $p(A)$, $p(C)$ the corresponding marginal probabilities Eq. (38) does not lead to a meaningful concept of independence.

Of course, one may always ask whether for sets A and C of a label space and a distribution p_i , $p_A(C)$ equals $p(C)$ and if so one may call A and C "independent." This seems, however, to be merely a formal analogy to the type of considerations which led us to (30) or to (34″). The ensuing multiplication of $p(A)$ and $p(C)$ does not correspond to a combination of

[2] We are conscious (see p. 4) of the difficulty of assigning, mathematically, a "dimension" to a discrete set of points. Nevertheless, it seems merely confusing to assign an r-dimensional *name* to the result of the tossing of *one* die, a name which is given by means of a vocabulary which has nothing to do with the actual experiment under observation.

[3] Examples of the inappropriate use of (38) as a definition of independence can be given if continuous label spaces are used, where the above-mentioned difficulty concerning "dimension" in the discrete case does not appear.

collectives and a statement like "event (2, 3, 4) is independent of event (1, 2) because (or if) $(p_2 + p_3 + p_4)(p_1 + p_2) = p_2$" does not convey the type of information which we connect in meaningful cases with the term independence.[4]

Our conclusion is that a definition of "independence" based on the measure equation (38) seems merely a formal generalization of a meaningful notion. It leads to inferences which have very little to do with the generally accepted and used meaning that the theory is supposed to reproduce. A development based on the notion of combinable collectives is appropriate for the definition of this fundamental and characteristic concept of probability theory.

APPENDIX ONE

THE CONSISTENCY OF THE NOTION OF THE COLLECTIVE. WALD'S RESULTS

Probability calculus as presented in this book is based on some concepts and ideas which may be briefly restated in a non-technical way, as follows:

(1) In probability calculus (or probability theory) we consider aggregates of uniform events, observations which can be repeated over and over, rather than isolated events; each observation leads to a result which can be expressed by a number (or by several numbers). As the conceptual counterpart of these observations and results, we introduce an infinite sequence $K = \{x_j\}$ of numbers representing the results or labels of the successive observations. For each label, a_i , $i = 1, 2, \ldots$ the limiting value of the relative frequency with which it occurs in K exists and is insensitive to place selections applied to the sequence.

(2) Such sequences are called collectives and the limiting frequency of a label a_i is the probability p_i of a_i in the collective K. The a_i together with the corresponding p_i form the probability distribution.

(3) By means of the repeated use of certain explicitly defined operations, probability distributions in new collectives are derived from given distributions in given collectives.

[4] Also, the *term* independence should be reserved for the repeatedly characterized meaningful situations as in the example of two dice, of eye color and hair color, of repeated draws.

Objections have been raised to the consistency of the concept of the collective. It has been said that proof is lacking that the concept is not empty and that it is free from contradictions; that the restriction stated by our concept of randomness [Section 4.2] is too severe, or not severe enough, etc. In past years, mainly 1925-1940, these questions were widely discussed and decisive results were obtained, probably unknown to many of today's students, who, often without knowing the frequency theory of the collective, follow the current fashion (cf. Appendix Two). In this appendix we shall present some basic mathematical results regarding the consistency of the notion of the collective. The problem arises in the simplest case, where the label set consists only of two labels, as well as in the most general cases. We shall take it up here for the case of a discrete label space which was the subject of Chapter I.

Wald[1] defines the collective in a way, which applies to any label space, as follows.

Let S be a label space, $\mathscr{S}$ a system of subsets of S, G a system of place selections, and $K = \{x_j\}$ an infinite sequence of labels. We call K a collective with respect to G and to $\mathscr{S}$, that is, $K = K(G, \mathscr{S})$ if:

(1) *For any element A of $\mathscr{S}$ there is a number $p(A)$ such that the relative frequency $n(A)/n$ of A in K converges with increasing n to $p(A)$.*

(2) *Any place selection belonging to G applied to K produces an infinite subsequence K' of K in which again for every A the relative frequency $n'(A)/n$ converges and the limit equals $p(A)$; $p(A)$ is then called the probability of A in K.*[2]

This probability is a non-negative, additive set function defined for all elements of $\mathscr{S}$; we call it briefly the probability distribution of K.

Wald poses the following problem: Let S be a label space. We should like to determine conditions holding for the system $\mathscr{S}$ of subsets of S, for the system G of place selections, and for the additive set function p, which assure us that a collective $K(G, \mathscr{S})$ exists whose probability distribution is p.

The answer, in the case of a discrete label space, is complete and will be easily understood by a reader of Chapter I:

Let S be a finite or infinite discrete label set $\{a_i\}$, $\mathscr{S}$ the system of all

[1] A. WALD, "Die Widerspruchsfreiheit des Kollektivbegriffs," *Ergeb. math. Kolloq.* 8 (1937), pp. 38-72. Our definition of the collective (p. 11) is influenced by Wald's analysis. For other literature, see footnote 9, p. 42.

[2] Reduced to a discrete label space $S = \{a_i\}$, this definition does not differ from ours (p. 12) although we introduce first the probabilities $p(a_i) = p_i$ of the single labels and form $p(A)$ by mixing.

subsets of S, G a system of denumerably many place selections, p a non-negative, σ-additive set function defined for all elements of $\mathscr{S}$ and such that

$$\sum_i p(a_i) \equiv \sum_i p_i = 1, \tag{a}$$

where the sum is over all (finitely many or countably many) labels a_i in S.

Then there exist infinitely many collectives $K(G, \mathscr{S})$ whose distribution equals p.[3]

This theorem states the existence of collectives for a discrete label space without any restriction regarding the set $\mathscr{S}$ and the only assumption regarding p is that $\Sigma_i p_i = 1$.[4] The restriction to a denumerable set of place selections is very weak.[5] We pointed out before that in no actual problem so far known do there arise more than countably many place selections[6] and that randomness may be considered equivalent to the assumption that *those* place selections G which occur in the solution of any particular problem do not change the limit(s) of relative frequencies in the respective collective(s) (Section 5.1). Wald also remarks that in the spirit of a formalized logic (for example, that of Russell and Whitehead) the number of all place selections that can be defined in words or symbols is countable. He calls a collective $K(G, \mathscr{S})$ *a Mises collective with respect to a formalized logic* R if G denotes the system of all place selections defined by mathematical rules belonging to R and he calls it *a Mises collective with respect to some problem* P if G denotes the set of all selections needed in the solution of P. Then, *under the conditions of the theorem* of the preceding page, *there exist Mises collectives with respect to any formalized logic R as well as with respect to any problem P, which needs only countably many place selections.*

It will be helpful to the reader if we indicate the basic lines of a proof of Wald's theorem[7] although we have to use some concepts which will be explicitly introduced in Chapter II only. Consider the simplest case of two labels 0 and 1. A collective is a sequence formed of zeros and ones and satisfying certain conditions. Now take a number x between zero and one and its binary expansion

[3] WALD, *loc. cit.*, p. 57.

[4] We remember that on pp. 18 ff. it was *proved* under assumption (a) that $p(A) = \Sigma_{a_i \in A} p(a_i)$ for any finite or infinite subset A of S and thus that $p(A)$ is a *completely* additive set function.

[5] In Wald's proof this restriction is sufficient but not necessary.

[6] Wald makes the stronger point that in any concrete problem the place selections used *are* at most denumerable.

[7] Wald's proof is quite complicated. Also his paper is hard to obtain. The simple proof given here is essentially due to E. Tornier.

$x = x_1/2 + x_2/2^2 + \cdots$, $x_i = 0, 1$.[8] The totality of these binary numbers maps one-to-one onto the points of the unit interval and we see that the totality of all infinite sequences $x_1, x_2, \ldots, x_i = 0, 1$ has measure one. Denoting by N_1 the number of one's among the first N digits of x, E. Borel has shown (1909) (see quotation, p. 7) *that the set of all x for which* $\lim_{N\to\infty} N_1/N = 1/2$ *has measure one.* This is well known and not difficult to prove. Analogously, the set of all decimal fractions which contain the "1" with asymptotic frequency 1/10 has measure one, etc., Hence, we see that the existence of a limit and even of the "correct" limit, 1/2, 1/10, ... is not a strange exception, but, so to speak, the rule.

How about place selections? Consider a fixed place selection s_1, as defined in Eqs. (5); the place selection defines a sequence $\alpha_1, \alpha_2, \ldots$ of those terms of the sequence $K = \{x\}$ which are "selected" and form the new sequence $K' = \{x'\}$. The rule is $x_{\alpha_1} = x_1'$, $x_{\alpha_2} = x_2'$, ... To every x-sequence corresponds one and only one x'-sequence. The converse is, of course not true: to a given x' correspond countably many x since all terms of the x-sequence except those with subscripts $\alpha_1, \alpha_2, \ldots$ remain undetermined. Consider a set A' of binary numbers x' on [0, 1] and denote by A the set of all the corresponding numbers x such that $x \in A$. Then one can prove without difficulty that if A' is measurable the same holds for A and $|A| = |A'|$. This rather elementary lemma will be proved in Chapter II, Sect. 4.5, as an application of the concept "basic set."

Now, consider the set U of all x in [0, 1]. According to Borel's theorem there exists in the x-space a set of sequences (binary numbers), U_0, such that for each of them $\lim_{N\to\infty} N_1/N$ converges and toward 1/2. Now apply to each x the given place selection s_1; each sequence x is transformed into a sequence x' and the whole set U is transformed into itself. The set U_0 of all sequences which converge toward 1/2 appears again in the x'-space; and there is a set U_1 in the x-space which corresponds to this U_0. All sequences x of U_1 have therefore the property that the selection s_1 transforms them into sequences with limiting frequencies 1/2 and $|U_1|$ equals one. In the same way, we consider countably many place selections $s_2, s_3, \ldots$ and obtain in the x-space sets U_ν, $\nu = 1, 2, 3, \ldots$ which are transformed by s_ν into the set U_0 of the x'-space, and for each ν, $|U_\nu| = 1$. The intersection K of the countably many sets $U_0, U_1, U_2, \ldots$ has also measure one (and therefore, *a fortiori* the cardinality of the continuum) and we have thus obtained K as a set of sequences of measure one each sequence having frequency limit 1/2, and being insensitive to the place selections s_ν; i.e., each sequence of K is a collective as postulated by v. Mises.

We considered here the simplest case with two labels and $p = q = 1/2$. It is, however, clear that the idea may be generalized. The generalization is immediate for k labels with rational probabilities $p_1, \ldots, p_k$ and almost immediate in the general case. (Use the procedure explained p. 68.)

It has thus been proved that the concept of the collective is neither inconsistent nor empty.[9] Further points will be discussed in Appendix

[8] For a binary rational we consider the unending expansion, e.g. 3/8 = .010111... rather than .011.

[9] We have reported here some of Wald's decisive results inasmuch as they concern the discrete case (his theorems I and II, pp. 45–46 and p. 57). His general result (theorems III, IV, p. 46 of his paper) concerns the label space which will be studied in Chapter II and will not be needed in our theory.

Two. A comparison of Mises' frequency theory with other foundations as well as a discussion of various viewpoints different from his own is contained in Chapter III of v. Mises' "Probability, Statistics, and Truth."[10]

APPENDIX TWO

MEASURE-THEORETICAL APPROACH VERSUS FREQUENCY APPROACH

We take it as understood that probability theory, like theoretical mechanics or geometry, is a scientific theory of a certain domain of observed phenomena. If we try to describe the known modes of scientific research we may say: all exact science starts with observations, which, at the outset, are formulated in ordinary language; these inexact formulations are made more precise and are finally replaced by axiomatic assumptions, which, at the same time, define the basic concepts. Tautological (= mathematical) transformations are then used in order to derive from these assumptions conclusions, which, after retranslation into common language, may be tested by observations, according to operational prescriptions.

Thus, there is in any sufficiently developed mathematical science a "middle part," a tautological or mathematical part, consisting of mathematical deductions. Nowadays, in the study of probability there is frequently a tendency to deal with this mathematical part in a careful and mathematically rigorous way, while little interest is given to the relation to the subject matter, to probability as a science.

Wald's investigations were preceded by important studies of Copeland, which, however, restrict the types of place selections admitted. Three of his papers are:

A. H. COPELAND, "The theory of probability from the point of view of admissible numbers," *Ann. Math. Statist.* **3** (1932), pp. 143–156; "Point set theory applied to the random selection of the digits of an admissible number," *Amer. J. Math.* **58** (1936), pp. 181–192; and, in particular, "Consistency of the conditions determining collectives," *Trans. Amer. Math. Soc.* **42** (1937), pp. 333-357.

Consistency of the collective is also proved by W. FELLER, "Über die Existenz von sogenannten Kollectiven," *Fund. Math.* **32** (1939), pp. 87–96. It has been said by Feller and others that the meaning of v. Mises had been changed by Wald's definitions. We wish to state here explicitly that v. Mises gratefully *endorsed* the precision added by Wald without feeling that his ideas had been altered.

[10] This is quoted in this book as "v. Mises [22]."

This is reflected in the fact that today the "measure-theoretical approach" is more generally favored than the "frequency approach" presented in this book. Cramér [5] very clearly expresses this point of view. "Following Kolmogorov,"[1] he says "we take as our starting point the observation that the probability $p(A)$ may be regarded as an additive set function of the set A. We shall, in fact, content ourselves by postulating mainly the existence of a function of this type defined for a *certain family* of sets A in the space to which our variable point X is restricted and such that $P(A)$ denotes the probability $X \in A$." And Halmos: "Probability is a branch of mathematics. Numerical probability is a measure function, that is, a finite, non-negative, and countably additive function P of elements in a Boolean σ-algebra B such that ... ."[2]

Now, such a description of the mathematical tools used in probability calculus seems to us only part of the story. Mass distributions, density distributions, and electric charge are likewise additive set functions. If there is nothing specific in probability, why do we define "independence" for probability distributions and not for mass distributions? Why do we consider random variables, convolutions, chains, and other specific concepts and problems of probability calculus?

The way we see it, probability is a highly mathematicized science, but it is not mathematics, just as hydrodynamics is not a branch of the theory of partial differential equations—although it suggests interesting and difficult problems in this field. Our aim in presenting probability theory as a mathematical science is to incorporate into the basic assumptions the idealized basic relations to reality, as this is done in mechanics and in other sciences. This approach has been criticized in the following comments [4]: "the von Mises definition involves a mixture of empirical and theoretical elements which is usually avoided in modern axiomatic theories. It would, for example, be comparable to defining a geometrical

---

[1] The starting point for Kolmogorov [17] (see our Chapter II, Section 3) is a set S or "elementary events." Subsets of S are called events. In the case of a discrete label space, the set $\mathcal{S}$ of *all* subsets of the discrete sample space S forms a Borel field, containing S (Section 7.3) the set on which we defined our probability distribution. In the general case, T is some field of subsets of S (the set of elementary events) and unless T is a Borel field the smallest Borel field T_σ over T is constructed by means of the Banach "extension theorem." On T_σ is defined a finite measure p, called probability, a real-valued non-negative countably additive set function such that $p(S) = 1$, etc. When speaking in this appendix of the measure theoretical approach or of the abstract or of the axiomatic approach we have, in general, Kolmogorov's setup in mind. Cramér, and many of his followers, take for T_σ the σ-field B_σ of the Borel sets in some Euclideans space. (Section 3).

[2] P. R. HALMOS, "The foundations of probability," *Amer. Math. Monthly* 51 (1944), pp. 493–510.

point as the limit of a chalk spot of infinitely decreasing dimensions." The "mixture of empirical and theoretical elements" is, in our opinion, unavoidable in a mathematical science. When in the theory of elasticity, we introduce the concepts of strain and stress, we cannot content ourselves by stating that these are symmetric tensors of second order. We have to bring in the basic assumptions of continuum mechanics, Hooke's law, etc., each of them a mixture of empirical and theoretical elements. Elasticity theory "is" not tensor analysis. Our definition of probability is comparable less to the "limit of a chalk spot" than to the definition of velocity as the limit of length per time or to the definition of specific density as the limit of mass per volume. Yet, all these definitions also have *something* of the chalk spot: the transition from observation to theoretical concepts cannot be completely mathematicized. It is not a logical conclusion but rather a choice, which, one believes, will stand up in the face of new observations.

Let us now state more specifically what we consider to be the essential features of the theory presented so far. First of all: it is a *frequency theory* of probability, in contrast to a "measure theory" of probability. Of course, in a "frequency theory" as well as in a "measure theory" the concepts of sets and their measures play an important role. Likewise, the concept of frequency appears in every measure theory of probability. However, by "frequency theory" or "frequency approach" we mean the following: *Whenever the term probability is used it relates to a (limit of a) frequency; and this frequency must be approximately verifiable, at least conceptually.* If we say that the probability of "6" for a die equals $\frac{1}{5}$, this implies that if we toss the die 10,000 times "6" will appear about 2000 times. The verification need not be so immediate but, directly, or indirectly by its consequences, some verification should be conceivable.[3] This is the view of v. Mises, Tornier, and Wald. On the other hand, if probability is defined as the measure of a set[4] and a relation to frequency is not incorporated in the theory but follows somewhere as an "obvious" interpretation, and with no necessary relation to verifiability, we speak of a *measure theory* of probability. As representatives we name Kolmogorov, Cramér, Fréchet, and, to some extent, Laplace.

We shall now explain two aspects of this distinction. The first one may be illustrated by a famous example (see Chapter IV, Section 4.2). Forming with the symbols 0 and 1 all combinations of n symbols we

[3] "Verification" is taken in the general sense of acceptance or rejection.

[4] This is an "*a priori* type" definition like Laplace's famous definition to which it reduces in simple cases; it has been aptly denoted by Fréchet as a "modernized classical" definition.

obtain 2^n different combinations (binary numbers). It can be shown that, for large enough n, the great majority of these 2^n numbers contains $n/2$ zeros and $n/2$ ones. More precisely: Jacob Bernoulli derived (1713) the following theorem: the larger the n, the larger the proportion of those binary numbers in which the relative numbers of zeros (or of ones) deviates from $\frac{1}{2}$ by less than a given ϵ. Obviously this is a purely arithmetic property of numbers (of binomial coefficients). But Bernoulli himself and most authors state the result in the following way: if one throws a "true" coin long enough, it is almost certain that the relative number of heads will deviate by less than ϵ from $\frac{1}{2}$. Certainly, this does not follow from the combinatorial theorem. The transition from that arithmetic theorem to a statement about "occurrence" can be justified only by defining a true coin (or any coin of probability p for "heads") in a way which *establishes a connection* between p and the frequency of occurrence of the event.[5]

The logical situation remains the same in the case of the sharper result expressed in the "strong" law of large numbers. We maintain the simple fact that if the deductive mathematical part of a theory is developed from axioms merely specifying the nature of the basic variables and functions involved, then relations to experience cannot be "derived mathematically." One cannot get out what one has not put in.

Great care must be given to incorporate the relation between theory and experience in an explicit and responsible way. This leads to our second comment. In measure theories of probability where all deductions are based on the measure definition, the relation to frequency is then often introduced as a more or less vague afterthought.[6] We illustrate the difference between this procedure and ours by the following example. One of the so-called Borel-Cantelli lemmas (see Chapter IV, Section 4.5) is often formulated as follows: "If the same trial with event-probability $p = \frac{1}{2}$ is repeated indefinitely, then the probability P that the event occurs finitely often only, tends to zero." Now, in a frequency theory a "probability that an event occurs only finitely often" has no place. Such a "probability" admits no verification, not even conceptually. We may perform any number of sequences of trials, each sequence of length n, but no matter how large the n, we can never decide whether any such segment is the beginning of an infinite sequence which contains finitely many successes only. The above "probability" P is the measure of a

[5] Actually, the opposite conclusion is often made: the above-quoted theorem (a purely arithmetical statement in the framework of measure theory) is considered to be the desired "bridge" between measure definition and experience.

[6] See for example many remarks in Kolmogorov [17], etc.

well-defined set and for Kolmogorov, this set, like all Borel sets, admits a probability. But in a frequency theory, such a measure cannot be considered a probability. The above-mentioned statement is a correct theorem of measure theory and may be quoted in this sence in any probability theory. *Our* aim is to build up a coherent theory of verifiable events. (More follows in Chapter II.)

From the works of Tornier, Wald, and Copeland (see quotations in Appendices One and Three) it appears that probability in a verifiable sense cannot be assigned to all measurable sets and not even to all Borel sets. The postulate worded by Cramér, that "any probability assigned to a specific event, must, in principle be liable to verification" is in contrast to axioms which assign probabilities to all Borel sets. In fact, if we are serious about the principle, instrumental for probability as a science, that an approximate verification should be conceivable for any probability statement then we arrive at the result (Chapter II, Section 5.6) that the sets of a well-defined field—quite different from that of Borel sets—and only these, should be assigned probabilities. We shall obtain this field F_1 by a very simple and direct extension of the field T of the collective studied in Chapter I. The sets of F_1 are characterized, mathematically, by the property that for each of them the measure of its "boundary" is zero. Such sets are said to "have content."[7] For measurable sets, which do not have content, the boundary has a positive measure, and we shall see that for such sets, *in principle*, no frequency interpretation, no verification is possible. We can use such sets and their measures freely in our deductions and conclusions, but, in the end, "probability" in contrast to measure should be restricted to situations which may be conceptually verified.[8] (In a sense, this is comparable to the auxiliary use of complex numbers in mechanics and electrotechnics.)

While we think that these features of frequency theories are essential, it is, perhaps, less important whether a theory is presented in an axiomatic form or not. One can certainly present our theory (Chapters I, II) in an axiomatic form. We preferred the approach presented here where,

[7] It is perhaps intuitively plausible that since the points of the boundary of a set A are arbitrarily close to its inner and outer points, to A and to A' = non-A, one cannot decide whether or not a point of the boundary satisfies a given condition. But if the measure of the boundary is zero this lack of decision does not influence the measure (=probability) of the set. Since in our theory (Chapter II), the "points," "sets of points," "boundary" have a specific meaning, this intuitive geometrical reasoning does not constitute a rigorous proof.

[8] It has been said v. Mises "defines" probability while Kolmogorov gives its mathematical theory. We think that our Chapters I and II show that v. Mises' frequency theory is a rigorous mathematical theory—different from that of Kolmogorov. (See also p. 110).

to quote one example, the fact that probability is a σ-additive set function *followed* from its frequency-explanation.

The reader might be reminded that our collective of Chapter I is already of considerable generality. Speaking of a collective, one thinks very often only of the simplest instances like the indefinitely continued throwing of a coin under unaltering circumstances. However (Section 7.3), in an n-dimensional collective the single "trial" may consist of the throwing of n "arbitrarily linked" "dice," with any finite or countable number of labels rather than six. In the corresponding collective K each single term is the result of such a complex trial (or observation) *which may include all forms of dependence.* Insensitivity toward place selection is postulated *within* K. The resulting distribution is $p(x_1, x_2, \ldots, x_n)$, $x_i = 1, 2, \ldots$; $i = 1, 2, \ldots, n$. The probability distribution in this n-dimensional collective has been proved to be a σ-additive set function over a σ-field. This is a description of our basic material. In Chapter II we shall extend this mathematical theory (with verifiability as the guiding principle) and shall obtain a mathematical theory of probability which is at the same time a scientific theory of probability.

We add, finally, that our setup seems of some pedagogical value. For the student of our theory it is natural to analyze each problem with regard to the collectives involved and the operations employed. Thus, the probabilistic side of a problem is stressed in contrast to the mere analytic side. Many apparent contradictions are resolved and difficulties removed by an analysis of the collectives and operations. It needs no effort to find in the literature on probability instances of conceptual mistakes juxtaposed with faultless mathematics.[9]

Today, we see at one extreme the "consumer of statistics" who wishes to apply ready-made statistics to his problem, be it in medicine, education, or linguistics. His desire is all to often to use statistical methods like prescriptions, recipes, whose application should require very limited statistical and even less probabilistic knowledge and understanding and a minimum of mathematics. At the other extreme, stand those mathematicians who are exclusively interested in the mathematical aspect of some problem, some part of the theory, and who therefore consider probability to be a branch of mathematics; who teach us that probability "is" measure theory, Boolean algebra, etc., and dismiss

[9] For the discussion of such instances see also H. GEIRINGER, "On the statistical investigation of transcendental numbers," in *Studies in Mathematics and Mechanics*, pp. 310–321 (section 3 and footnote 4), New York, 1954. Of course, the collective, etc., is not *needed*, in order to avoid conceptual mistakes. It is, however, dangerous, if an inexperienced student gets the impression that the probabilistic and statistical side of a concept or problem is simple and trivial for one who knows measure theory.

frequency theory as "awkward mathematically." Between the extremes, the conception of probability theory as a mathematical science leads to a frequency theory of probability, much in need of the mathematicians' ideas and ingenuity, but free of a confusion of task and tool.