FIRST PART

PRINCIPLES

CHAPTER I

INTRODUCTORY REMARKS

1. In the most varied fields of practical and scientific experience, cases occur where certain observations or trials may be repeated a large number of times under similar circumstances. Our attention is then directed to a certain quantity, which may assume different numerical values at successive observations. In many cases each observation yields not only one, but a certain number of quantities, say k, so that generally we may say that the result of each observation is a definite point X in a space of k dimensions ($k \geqslant 1$), while the result of the whole series of observations is a sequence of points: $X_1, X_2, \ldots$.

Thus if we make a series of throws with a given number of dice, we may observe the sum of the points obtained at each throw. We are then concerned with a variable quantity, which may assume every integral value between m and $6m$ (both limits inclusive), where m is the number of dice. On the other hand, in a series of measurements of the state of some physical system, or of the size of certain organs in a number of individuals belonging to the same biological species, each observation furnishes a certain number of numerical values, i.e. a definite point X in a space R of a fixed number of dimensions.

In certain cases, the observed characteristic is only indirectly expressed as a number. Thus if, in a mortality investigation, we observe during one year a large number of persons, we may at each observation (i.e. for each person) note the *number of deaths* which take place during the year, so that in this case the observed

quantity assumes the value 0 or 1 according as the corresponding person is alive at the end of the year or not.

In a given class of observations, let R denote the set of points which are *a priori* possible positions of our variable point X, and let S be a sub-set of R. Further, let a series of n observations be made, and count the number ν of those observations, where the following *event* takes place: *the point X determined by the observation belongs to S*. Then the ratio ν/n is called the *frequency* of that event or, as we may shortly put it, *the frequency of the relation* (*or event*) $X \subset S$. Obviously any such frequency always lies between 0 and 1, both limits inclusive. If $S = S_1 + S_2$, where S_1 and S_2 have no common point, and if ν_1/n and ν_2/n are the frequencies corresponding to S_1 and S_2, we obviously have $\nu = \nu_1 + \nu_2$ and thus

$$\nu/n = \nu_1/n + \nu_2/n. \tag{1}$$

When we are dealing with such frequencies, a certain peculiar kind of regularity very often presents itself. This regularity may be roughly described by saying that, for any given sub-set S, the frequency of the relation (or event) $X \subset S$ *tends to become more or less constant as n increases*. In certain cases, such as e.g. cases of biological measurements, our observations may be regarded as samples from a very large or even infinite population, so that for indefinitely increasing n the frequency would ultimately reach an ideal value, characteristic of the total population.

It is thus suggested that in cases where the above-mentioned type of regularity appears, we should try to introduce a number $P(S)$ to represent such an ideal value of the frequency ν/n corresponding to the sub-set S. The number $P(S)$ is then called *the probability of the sub-set S, or of the event $X \subset S$*. It follows from (1) that we should obviously choose $P(S)$ such that

$$P(S_1 + S_2) = P(S_1) + P(S_2) \tag{2}$$

for any two sub-sets S_1 and S_2 of R which have no common point. Further, it is obvious that we should always have $P(S) \geqq 0$ and that for the particular set $S = R$ we should have $P(R) = 1$.

The investigation of *set functions* of the type $P(S)$ and their mutual relations is the object of the Mathematical Theory of Probability. This theory should be considered as a branch of Pure Mathematics, founded on an axiomatic basis, in the same sense as Geometry or Theoretical Mechanics.[1] Once the fundamental conceptions have been introduced and the axioms have been laid down (and in this procedure we are, of course, guided by empirical considerations), the whole body of the theory should be constructed by purely mathematical deductions from the axioms. The practical value of the theory will then have to be tested by experience, just in the same way as a theorem in euclidean geometry, which is intrinsically a purely mathematical proposition, obtains a practical value because experience shows that euclidean geometry really conforms with sufficient accuracy to a large group of empirical facts.

We finally point out that, in order to build a perfectly general mathematical theory of the phenomena encountered in connection with experimental situations of the type considered here, it would be necessary to remove the restriction that R should be a space of a finite number of dimensions. We should then have to regard X as an observed point in some space R of a more general nature. For the purposes of this book we shall, however, restrict ourselves to the case when R has a finite—although in some cases very large—number of dimensions.

2. The axiomatic basis of a theory may, of course, always be constructed in many different ways, and it is well known that, with respect to the foundations of the Theory of Probability, there has been a great diversity of opinions.

The type of statistical regularity indicated above was first observed in connection with ordinary games of chance with cards, dice, etc., and this gave occasion to the origin and early development of the theory.[2] In every game of this character, all

[1] This view seems to have been first explicitly expressed by v. Mises [**2**].
[2] Cf. Todhunter [**1**].

the results that are *a priori* possible may be arranged in a finite number of cases which are supposed to be perfectly symmetrical. This led to the famous *principle of equally possible cases* which, after having been more or less tacitly assumed by earlier writers, was explicitly framed by Laplace [**1**], as the fundamental principle of the whole theory.

During the subsequent discussion of this principle, it has been maintained by various authors that the validity of the principle of equally possible cases is necessarily restricted to the field of games of chance. Attempts have been made[1] to establish the theory on an essentially different basis, the probabilities being directly defined as ideal values of statistical frequencies. The most successful attempt on this line is due to v. Mises [**2**, **3**], who endeavours to reach in this way an axiomatic foundation of the theory in the modern sense.

The fundamental conception of the v. Mises theory is that of a "*Kollektiv*", by which is meant an unlimited sequence K of similar observations, each furnishing a definite point belonging to an *a priori* given space R of a finite number of dimensions. The first axiom of v. Mises then postulates the existence of the limit

$$\lim_{n\to\infty} \nu/n = P(S) \tag{3}$$

for every simple sub-set $S \subset R$, while the second axiom requires that the analogous limit should still exist and have the same value $P(S)$ for every sub-sequence K' that can be formed from K according to a rule such that it can always be decided whether the nth observation of K should belong to K' or not, *without knowing the result of this particular observation*. It does, however, seem difficult to give a precise mathematical meaning to the condition printed in italics, and the attempts to express the second axiom in a more rigorous way do not, so far, seem to have reached satisfactory and easily applicable results. Though fully recognizing the value of a system of axioms based on the pro-

[1] For the history of these attempts, cf. Keynes [**1**], chaps. VII–VIII.

perties of statistical frequencies, I think that these difficulties must be considered sufficiently grave to justify, at least for the time being, the choice of a different system.

The underlying idea of the system that will be adopted here may be roughly described in the following simple way: *The probability of an event is a definite number associated with that event; and our axioms have to express the fundamental rules for operations with such numbers.*

Following Kolmogoroff [4], we take as our starting-point the observation made above (cf. (2)) that the probability $P(S)$ may be regarded as an *additive function of the set* S. We shall, in fact, content ourselves by postulating mainly the existence of a function of this type, defined for a certain family of sets S in the k-dimensional space R_k to which our variable point X is restricted. and such that $P(S)$ denotes the probability of the relation $X \subset S$.

Thus the question of the validity of the relation (3) will not at all enter into the mathematical theory. For the *empirical verification* of the theory it will, on the other hand, become a matter of fundamental importance to know if, in a given case, (3) is satisfied with a practically sufficient approximation. Questions of verification and application fall, however, outside the scope of the present work, which will be exclusively concerned with the development of the purely mathematical part of the subject.

3. Before giving the explicit statement of our axioms, it will be convenient to discuss here a few preliminary questions related to the theory of point sets and (generalized) Stieltjes integrals in spaces of a finite number of dimensions.[1]

In the first place, we must define the family F of sets S, for which we shall want our additive set function $P(S)$ to be given. If $X = (\xi_1, \ldots, \xi_k)$ belongs to the k-dimensional euclidean space

[1] Reference may be made to the treatises by Hobson [1], Lebesgue [1] and de la Vallée Poussin [1].

R_k, the family F should obviously contain every k-dimensional interval J defined by inequalities of the form

$$a_i < \xi_i \leqq b_i \quad (i = 1, 2, \ldots, k),$$

as we may always want to know the probability of the relation $X \subset J$. It is also obvious that F should contain every set S constructed by performing on intervals J a finite number of additions, subtractions and multiplications. It is even natural to require that it should be possible to perform these operations an infinite number of times without ever arriving at a set S such that the value of $P(S)$ is not defined. Accordingly, we shall assume that $P(S)$ is defined for all *Borel sets*[1] S of R_k.

Every set which can be constructed from intervals J by applying a finite or infinite number of times the three elementary operations is a Borel set. If $S_1, S_2, \ldots$ are Borel sets in R_k, this also holds true for the two sets

$$\limsup S_n = \lim (S_n + S_{n+1} + \ldots),$$

$$\lim \inf S_n = \lim (S_n S_{n+1} \ldots).$$

If $\limsup S_n$ and $\lim\inf S_n$ are identical, we put

$$\lim S_n = \limsup S_n = \lim\inf S_n,$$

and thus $\lim S_n$ is also a Borel set. In particular, the sum and product of an infinite sequence of Borel sets are always Borel sets.

If no two of the sets S_i have a common point, it follows from the additive property (2) that

$$P(S_1 + \ldots + S_n) = P(S_1) + \ldots + P(S_n)$$

for every finite n. Since the limit $S_1 + S_2 + \ldots$ always exists and is a Borel set, it is natural to require that this relation should hold even as $n \to \infty$, so that we should have

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

A set function with this property will be called *completely additive*, and it will be assumed that the function $P(S)$ is of this type.

[1] Cf. Hobson [1], I, p. 179; Lebesgue [1], p. 117; de la Vallée Poussin [1], p. 33.

Consider now a real-valued point function $g(X)$, defined for all points $X=(\xi_1,\ldots,\xi_k)$ in R_k. $g(X)$ is said to be *measurable* B[1] if, for all real a and b, the set of points X such that $a<g(X)\leqq b$ is a Borel set. Similarly, a vector function $Y=f(X)$, where $Y=(\eta_1,\ldots,\eta_\mathfrak{k})$ belongs to a certain $\mathfrak{k}$-dimensional space $\mathfrak{R}_\mathfrak{k}$, is measurable B if every component η_i, regarded as a function of X, is measurable B. If $\mathfrak{S}$ denotes any Borel set in $\mathfrak{R}_\mathfrak{k}$, and if S is the set of all points X in R_k such that $f(X)\subset\mathfrak{S}$, then S is also a Borel set. (If $f(X)$ never assumes a value belonging to $\mathfrak{S}$, S is of course the empty set.) If $f_1, f_2, \ldots$ are measurable B, so are $f_1\pm f_2$, $f_1 f_2$, f_1^{-1}, $\limsup f_n$, $\liminf f_n$ and, in the case of convergence, $\lim f_n$.

All sets of points with which we shall have to deal in the sequel are Borel sets, while all point functions are measurable B. Generally this will not be explicitly mentioned, and should then always be tacitly understood.

A *Lebesgue-Stieltjes integral* with respect to the completely additive set function $P(S)$ is, for every bounded and non-negative $g(X)$ and for every set S, uniquely defined by the postulates

$$\text{(A)}\qquad \int_{S_1+S_2} g\,dP=\int_{S_1} g\,dP+\int_{S_2} g\,dP,$$

S_1 and S_2 having no common point, and

$$\text{(B)}\qquad \int_S (g_1+g_2)\,dP=\int_S g_1\,dP+\int_S g_2\,dP,$$

$$\text{(C)}\qquad \int_S g\,dP\geqq 0,$$

$$\text{(D)}\qquad \int_S 1\,.\,dP=P(S).$$

If g is not bounded, we put $g_M=\min(g,M)$ and define $\int_S g\,dP$ as the limit of $\int_S g_M\,dP$ as $M\to\infty$. If the limit is finite, g is said

[1] Cf. Hobson [1], I, p. 563; de la Vallée Poussin [1], p. 34.

to be integrable over S with respect to $P(S)$. The extension to functions g which are not of constant sign is performed by putting

$$2\int_S g\,dP = \int_S (|g|+g)\,dP - \int_S (|g|-g)\,dP.$$

For any g such that $|g| < C$ throughout the set S, we then have the mean value theorem

$$\left|\int_S g\,dP\right| < C\,P(S).$$

Let $g_1, g_2, \ldots$ be a sequence of functions such that for all points of S we have $|g_n| < g$, where g is integrable. Then if $\lim g_n$ exists for every point of S, except possibly for a certain set of points $S_1 \subset S$ such that $P(S_1) = 0$, we have

$$\lim \int_S g_n\,dP = \int_S \lim g_n\,dP.$$

It follows that the theorems on continuity, differentiation and integration with respect to a parameter, etc. which are known from elementary integration theory extend themselves immediately to integrals of the type $\int_S g(X,t)\,dP$, where t is a parameter.

The ordinary theorems on repeated integrals[1] are also easily extended to integrals of the type here considered. In particular we have the following result which will be used in Chapter III. Let $P(S)$ be defined in a two-dimensional space R_2 and such that for every two-dimensional interval J $(a_1 < \xi_1 \leqq b_1, a_2 < \xi_2 \leqq b_2)$ we have

$$P(J) = P_1(J_1)\,P_2(J_2),$$

where $P_1(S)$ and $P_2(S)$ are completely additive set functions in R_1 while J_i denotes the one-dimensional interval $a_i < \xi_i \leqq b_i$. Then if the function $g_1(\xi_1)\,g_2(\xi_2)$ is integrable over R_2 with respect to $P(S)$, we have

$$\int_{R_2} g_1(\xi_1)\,g_2(\xi_2)\,dP = \int_{R_1} g_1(\xi_1)\,dP_1 \int_{R_1} g_2(\xi_2)\,dP_2.$$

[1] Cf. Hobson [1], I, p. 626; de la Vallée Poussin [1], p. 50.

CHAPTER II

AXIOMS AND PRELIMINARY THEOREMS

1. We now proceed to the explicit statement of our axioms.[1] In accordance with the preceding chapter, we denote by R_k a k-dimensional euclidean space with the variable point $X=(\xi_1, \ldots, \xi_k)$, and we consider the family of all Borel sets S in R_k.

Axiom 1. *To every S corresponds a non-negative number $P(S)$, which is called the probability of the relation (or event) $X \subset S$.*

Axiom 2. *We have $P(R_k)=1$.*

Axiom 3. *$P(S)$ is a completely additive set function, i.e. we have*

$$P(S_1+S_2+\ldots)=P(S_1)+P(S_2)+\ldots,$$

where $S_1, S_2, \ldots$ are Borel sets, no two of which have a common point.

The variable point X is then called a *random variable* (or random point, random vector). The set function $P(S)$ is called the *probability function* of X, and is said to define the *probability distribution* in R_k which is attached to the variable X. It is often convenient to use a concrete interpretation of a probability distribution as a distribution of mass of the total amount 1 over R_k, the quantity of mass allotted to any Borel set S being equal to $P(S)$.

It follows immediately from the axioms that we always have

$$0 \leqq P(S) \leqq 1,$$

and

$$P(S)+P(S^*)=1,$$

where S and S^* are complementary sets. Further, if S_1 and S_2 are two sets such that $S_1 \supset S_2$, we have $S_1=S_2+(S_1-S_2)$ and thus

$$P(S_1) \geqq P(S_2). \tag{4}$$

[1] The fact that we restrict ourselves here to Borel sets in R_k permits some formal simplification of the system of axioms given by Kolmogoroff [**4**], and of the immediate conclusions drawn from the axioms.

Theorem 1. *For any sequence of Borel sets* $S_1, S_2, \ldots$ *in* R_k, *we have*

$$P(\limsup S_n) \geqq \limsup P(S_n),$$
$$P(\liminf S_n) \leqq \liminf P(S_n).$$

Hence, if $\lim S_n$ *exists, so does* $\lim P(S_n)$, *and we have*

$$P(\lim S_n) = \lim P(S_n). \tag{5}$$

In order to prove this theorem, we shall first show that (5) holds for any *monotone* sequence $\{S_n\}$. If $\{S_n\}$ is an *increasing* sequence, we may in fact write

$$\lim S_n = S_1 + (S_2 - S_1) + (S_3 - S_2) + \ldots,$$

and thus obtain from Axiom 3

$$\begin{aligned} P(\lim S_n) &= P(S_1) + P(S_2 - S_1) + P(S_3 - S_2) + \ldots \\ &= P(S_1) + (P(S_2) - P(S_1)) + (P(S_3) - P(S_2)) + \ldots \\ &= \lim P(S_n). \end{aligned}$$

For a *decreasing* sequence $\{S_n\}$, the same thing is shown by considering the increasing sequence formed by the complementary sets S_n^*.

For any sequence $\{S_n\}$, whether monotone or not, we have (cf. I, § 3) $\limsup S_n = \lim (S_n + S_{n+1} + \ldots)$. Now, $S_n + S_{n+1} + \ldots$ is obviously the general element of a decreasing sequence, so that

$$P(\limsup S_n) = \lim P(S_n + S_{n+1} + \ldots). \tag{6}$$

For every $r = 0, 1, \ldots$, we have $S_n + S_{n+1} + \ldots \supset S_{n+r}$, and thus by (4)

$$P(S_n + S_{n+1} + \ldots) \geqq P(S_{n+r}),$$
$$P(S_n + S_{n+1} + \ldots) \geqq \limsup P(S_n).$$

We thus obtain from (6)

$$P(\limsup S_n) \geqq \limsup P(S_n).$$

Hence the inequality for $P(\liminf S_n)$ is obtained by considering the sequence $\{S_n^*\}$ of complementary sets and using the identity $\liminf S_n = (\limsup S_n^*)^*$. Thus Theorem 1 is proved.

In the particular case when every point X of R_k belongs at most to a finite number of the sets S_n, $\lim S_n$ is the empty set, and it follows that we have $\lim P(S_n) = 0$.